PORTFOLIO MANAGEMENT RESEARCH

SPRING 2020 VOLUME 2 NUMBER 2

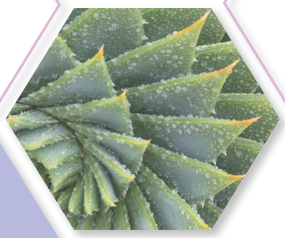
the journal of financial data science

JFDS.pm-research.com

"expanding the frontiers of data-driven investing

—FRANK J. FABOZZI, MARCOS LÓPEZ DE PRADO, AND JOSEPH SIMONIAN

PORTFOLIO
MANAGEMENT
RESEARCH



THE JOURNAL OF impact&esg INVESTING

With environmental, social, and governance factors becoming increasingly important when making investment decisions and managing risk, our new journal will ensure you have all the research you need to manage and enhance your portfolios effectively.

launching summer **2020**

JESG.pm-research.com

the journal of financial data science

VOLUME 2, NUMBER 2 SPRING 2020

Frank J. Fabozzi	Editor
Marcos López de Prado	Editor
Joseph Simonian	Editor
Francesco A. Fabozzi	Managing Editor
Mitchell Gang	Production Editor
Deborah Brouwer	Production and Design Manager
Mark Adelson	Content Director
Rosie Instance	Marketing Manager
William Law	Account Manager
Nikol Madjarova	Account Manager
Ryan C. Meyers	Account Manager
David Rowe	Reprints Manager
Mark Lee	Advertising Director

In the lead article, "It's All About Data: *How to Make Good Decisions in a World Awash with Information*," Mehrzad Mahdavi and Hossein Kazemi explain how, in the financial sector, the rise of big and alternative data has resulted in significant new opportunities in all areas of the financial industry, including risk management, portfolio construction, investment banking, and insurance. To build trusted outcomes in artificial intelligence/machine learning initiatives, the roles of financial professionals are critical. Because there are many nuances in using big data, the authors explain why there is a need for vetted protocols and methods in the selection of data sets and algorithms. There are best practices and guidelines in effectively reducing the risks of using artificial intelligence and machine learning, including overfitting, lack of interpretability, biased inputs, and unethical use of data. Mahdavi and Kazemi emphasize that a key success factor for the deployment of artificial intelligence initiatives is talent and bridging the skills gap. Given the major shortage of talent in this space, practical training of employees and continued education are key to ensuring sustainable growth in the fast-moving field of trusted data-driven decision making.

Machine learning can help asset managers in most portfolio construction tasks, such as idea generation, alpha factor design, asset allocation, weight optimization, position sizing, and the testing of strategies. Derek Snow's article, "Machine Learning in Asset Management—*Part 2: Portfolio Construction—Weight Optimization*," is the second in a series of articles dealing with machine learning in asset management focusing on weight optimization strategies. Specifically, Snow explains the different weight optimization methods that can be used for supervised, unsupervised, and reinforcement learning. Within the supervised learning framework, three techniques are explained. The first is a more traditional linear approach using OLS, Ridge, and LASSO regressions. The second is a nonlinear deep learning approach using autoencoders. The third is a Bayesian sentiment method. Three methods for unsupervised methods are principal component analysis, hierarchical clustering analysis, and network graphs. Finally, Snow explains portfolio weight optimization for the deep reinforcement learning approach.

Zihao Zhang, Stefan Zohren, and Stephen Roberts adopt deep reinforcement learning algorithms to design trading strategies for continuous futures contracts in their article, "Deep Reinforcement Learning for Trading." They begin by discussing the connection between modern portfolio theory and the reinforcement learning reward hypothesis, showing that they are equivalent if a linear utility function is used. Both discrete and continuous action spaces are considered, and volatility scaling is incorporated to create reward functions that scale trade

positions based on market volatility. They backtest their proposed algorithms on 50 very liquid futures contracts for the period 2011 to 2019 and investigate how performance varies across different asset classes (commodities, equity indexes, fixed income, and foreign exchange markets). When their algorithms are compared to classical time-series momentum strategies, Zhang, Zohren, and Roberts find that their method outperforms baseline models, delivering positive profits despite heavy transaction costs. The authors conclude that their proposed algorithms can follow large market trends without changing positions and can also scale down, or hold, through consolidation periods.

Quantitative asset management firms typically rely on a multitude of investment signals to formulate models to drive alpha. Ideally, each signal will provide information that is both predictively useful and complementary to the other signals feeding into an investment process. For this reason, a critical component of quantitative investing is model validation. A balancing of informative power and simplicity is required in the construction of any model employed by quantitative asset management firms. In "Modular Machine Learning for Model Validation: *An Application to the Fundamental Law of Active Management*," Joseph Simonian introduces an approach to model validation that he calls "modular machine learning" and applies the model to build a methodology that can be used to evaluate investment signals within the conceptual scheme provided by Grinold and Kahn's fundamental law of active management. The author's framework is modular in two respects: (1) it is composed of independent computational components, each using the output of another as its input and (2) it is characterized by the distinct role played by traditional econometric and data science methodologies. An econometric model in the first module is used to classify data in an economically intuitive way. To demonstrate how modular machine learning works, Simonian provides an example using the well-known Fama–French factors as test signals.

In "Interactions and Interconnectedness Shape Financial Market Research," Otto Loistl and Gueorgui S. Konstantinov investigate two issues relevant to financial data science. The first issue covers the entire production chain "from orders to prices" by modeling the microstructure of stock exchanges. The authors explain how data-driven research can model decisions to place orders and to generate prices by matching orders accordingly. The second issue relates to the price interconnectedness of markets by networks. Network analysis for factors and assets and centrality-based solutions in portfolio management are advantageous for capturing nonlinearity in data. The authors show that interactions shape a market's performance and provide two examples that market participants interact, learn, and trade.

Financial modeling and analysis have traditionally relied primarily on financial econometrics, which applies statistical methods to the problems of finance. This has changed in recent years as advances in analytical and algorithmic methods, powerful technological solutions, and the expansion of the financial data ecosystem now offer novel and robust ways to solve existing and new problems in finance. Tamer Khraisha, in "A Holistic Approach to Financial Data Science: *Data, Technology, and Analytics*," discusses in detail the financial data ecosystem, illustrating the types and structures of financial data and their distinguishing features. The author proposes a comprehensive, financial-domain–oriented approach to financial data science management that can be used by researchers and practitioners to exploit the opportunities arising from the evolving finance landscape.

Principal component analysis (PCA) is a useful tool when trying to construct factor models from historical asset returns. Using PCA, Marco Avellaneda, Brian Healy, Andrew Papanicolaou, and George Papanicolaou demonstrate that a relatively small number of factors can account for most of the variation in the collective movements of the implied volatilities derived from US equity options. In their article, "PCA for Implied Volatility Surfaces," they show that this market factor is the index resulting from the daily compounding of a weighted average of implied-volatility returns, with weights based on the options' open interest and Vega. Based on their analysis of the singular vectors derived from the tensor structure of the implied volatilities of S&P 500 constituents, the authors also find evidence that some type of open interest- and Vega-weighted index should be one of at least two significant factors in this market.

Asset management and custody banks are the two primary subindustries within the capital market industry. In their article "Evaluating Cost Efficiencies in Asset Management and Custody Banks Using Data Envelopment Analysis," D. K. Malhotra, Rashmi Malhotra, and Robert Nydick use a data envelopment analysis (DEA) model to evaluate the economies of scale in the asset management and custody bank subindustry for the period of 2010 to 2017. Measuring company size using a firm's total assets under management and total revenue generated, the authors evaluate cost efficiencies with cost measured in terms of the two important components of total expenses (total distribution expenses and general administrative and selling expenses). Using a returns-to-scale DEA model, they find that whenever asset management and custody banks reach their most productive scale size and begin to experience decreasing returns to scale, they make structural changes by adopting new technologies and/or changes in the processes, so that their total operating expenses decline and they can grow again without a proportionate increase in operating costs.

Francesco A. Fabozzi
Managing Editor

the journal of financial data science

Editors

Frank J. Fabozzi
EDHEC Business School

Marcos López de Prado
True Positive Technologies
and Cornell University

Joseph Simonian
Autonomous Investment
Technologies

Managing Editor

Francesco A. Fabozzi
Frank J. Fabozzi Associates

Advisory Board

Irene Aldridge
Cornell University and AbleMarkets

Robert Arnott
Research Affiliates

Joseph Cerniglia
University of Pennsylvania

Rama Cont
Oxford University

Germán G. Creamer
Stevens Institute of Technology

Sanjiv Ranjan Das
Santa Clara University

Campbell Harvey
Fuqua School of Business
Duke University

Michael Imerman
Peter F. Drucker and Masatoshi Ito
Graduate School of Management
Claremont Graduate University

Bruce Jacobs
Jacobs Levy Equity Management

Ron Kahn
BlackRock Financial Management

Kathryn Kaminski
AlphaSimplex Group LLC

Hossein Kazemi
CISDM/Isenberg School
of Management and
University of Massachusetts at Amherst

Petter Kolm
Courant Institute,
New York University

Han Liu
Northwestern University

Anthony Ledford
Man-AHL

Yongjae Lee
Ulsan National Institute
of Science and Technology

Chen Liang
Bridgewater Associates

Jim Kyung-Soo Liew
Johns Hopkins University

Ganesh Mani
Carnegie Mellon

Dhagash Mehta
The Vanguard Group

Paul F. Mende
MIT Sloan School of Management

George Mussalli
PanAgora

Warren Pennington
Vanguard

Sidney C. Porter
FEV Analytics

Arun Verma
Bloomberg

Zhipeng (Alan) Yan
Martin Tuchman School of Management,
New Jersey Institute
of Technology

The Journal of Financial Data Science®, (ISSN# 2640-3943) is published quarterly by Pageant Media, Ltd., 41 Madison Avenue, 20th Floor, New York, NY 10010. Phone (646) 891-2110. Copyright © 2020 Pageant Media, Ltd. POSTMASTER: Send address changes for *The Journal of Financial Data Science* to Pageant Media, Ltd., 41 Madison Avenue, 20th Floor, New York, NY 10010 USA, Att: Journals Department. Reproduction in whole or in part without written permission is prohibited.

Manuscript: *The Journal of Financial Data Science* considers but assumes no responsibility for unsolicited manuscripts. Unsolicited manuscripts will not be returned. Instructions for submission of manuscripts are printed at the end of the journal.

Subscriptions: For subscription enquiries please contact a member of the team at +1 646 931 9045/+44 20 7139 1600 or at pm-research@pageantmedia.com. Printed by The Manson Group, St Albans, UK.

The opinions expressed are those of the individual authors. Portfolio Management Research is not responsible for the accuracy, completeness, or timeliness of the information contained in the articles herein. If you require expert advice, you should seek the services of a competent professional. No statement in this journal is to be construed as a recommendation to buy or sell securities. Product names mentioned in the articles in this journal may be trademarks or service marks of their respective owners.

It's All About Data: *How to Make Good Decisions in a World Awash with Information* 8

Mehrzad Mahdavi and Hossein Kazemi

The rise of big and alternative data has created significant new business opportunities in the financial sector. As we start on this journey of fast-moving technology disruption, financial professionals have a rare opportunity to balance the exponential growth of artificial intelligence (AI)/data science with ethics, bias, and privacy to create trusted data-driven decision making. In this article, the authors discuss the nuances of big data sets that are critical when one considers standards, processes, best practices, and modeling algorithms for the deployment of AI systems. In addition, this industry is widely guided by a fiduciary standard that puts the interests of the client above all else. It is therefore critical to have a thorough understanding of the limitations of our knowledge, because there are many *known unknowns* and *unknown unknowns* that can have a significant impact on outcomes. The authors emphasize key success factors for the deployment of AI initiatives: talent and bridging the skills gap. To achieve a lasting impact of big data initiatives, multidisciplinary teams with well-defined roles need to be established with continuing training and education. The prize is the finance of the future.

TOPICS: ***Simulations, big data/machine learning****

Machine Learning in Asset Management—*Part 2: Portfolio Construction—Weight Optimization* 17

Derek Snow

This is the second in a series of articles dealing with machine learning in asset management. This article focuses on portfolio weighting using machine learning. Following from the previous article (Snow 2020), which looked at trading strategies, this article identifies different weight optimization methods for supervised, unsupervised, and reinforcement learning frameworks. In total, seven submethods are summarized, with the code made available for further exploration.

TOPICS: ***Big data/machine learning, analysis of individual factors/risk premia, portfolio construction, performance measurement****

Deep Reinforcement Learning for Trading 25

Zihao Zhang, Stefan Zohren, and Stephen Roberts

In this article, the authors adopt deep reinforcement learning algorithms to design trading strategies for continuous futures contracts. Both discrete and continuous action spaces are considered, and volatility scaling is incorporated to create reward functions that scale trade positions based on market volatility. They test their algorithms on 50 very liquid futures contracts from 2011 to 2019 and investigate how performance varies across different asset classes, including commodities, equity indexes, fixed income, and foreign exchange markets. They compare their algorithms against classical time-series momentum strategies and show that their method outperforms such baseline models, delivering positive profits despite heavy transaction costs. The experiments show that the proposed algorithms can follow large market trends without changing positions and can also scale down, or hold, through consolidation periods.

TOPICS: ***Futures and forward contracts, exchanges/markets/clearinghouses, statistical methods, simulations****

Modular Machine Learning for Model Validation: *An Application to the Fundamental Law of Active Management* 41

Joseph Simonian

The author introduces a modular machine learning framework for model validation in which the output

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

from one procedure serves as the input to another procedure within a single validation framework. A defining feature of the described methodology is the use of both traditional econometrics and data science. The author uses an econometric model in the first module to classify data in an economically intuitive way. Proceeding modules apply data science techniques to evaluate the predictive characteristics of the model components. The author applies his framework to the fundamental law of active management, a well-known formal characterization of portfolio managers' alpha generation process. In contrast to standard applications of the law, in which it has been used to evaluate a manager's existing active management process, the author recasts the law within his framework as a means to test investment signals for potential use, individually or collectively, in a manager's investment process. To illustrate how this application works, the author provides an example using the well-known Fama–French factors as test signals.

TOPICS: ***Statistical methods, simulations, big data/machine learning****

Interactions and Interconnectedness Shape Financial Market Research 51

Otto Loistl and Gueorgui S. Konstantinov

In this article, the authors investigate two fields that might be relevant for financial data sciences. The first issue covers the entire production chain "from orders to prices" by realistically modeling stock exchanges' microstructure (e.g., NASDAQ and Xetra). Specifically, the authors show how data-driven research can model decisions to place orders and to generate prices by matching orders accordingly. The other issue is price interconnectedness at markets by networks. The authors show that interactions shape a market's performance. Emergence comprises the interactions at markets; as such, the collective may not be equal to the sum of individual activities. As a consequence, the assumption that markets are in equilibrium and that arbitrage opportunities do not exist can be replaced by more realistic working hypotheses. The authors show with the two examples that market participants interact, learn, and trade. These individual interactions can be described as organized complexity. Whereas calculus may not support explicit modeling of interactions, the age of big data permits their modeling and application of innovative concepts, such as network solutions for asset allocation, which can be modeled using machine learning. This article illustrates that assertion with concrete examples.

TOPICS: ***Equity portfolio management, statistical methods, simulations, big data/machine learning****

A Holistic Approach to Financial Data Science: *Data, Technology, and Analytics* 64

Tamer Khraisha

The scientific analysis of financial data, both for practical and theoretical purposes, has continuously been an active and evolving area. Traditionally, financial modeling and analysis have primarily relied on financial econometrics, which applies statistical methods to the problems of finance. Recently, advances in analytical and algorithmic methods, powerful technological solutions, and the expansion of the financial data ecosystem have created new opportunities and provided novel and robust ways to solve existing and new problems in finance. To expand the scope of financial econometrics and incorporate these novelties, the field of financial data science (FDS) has emerged. Crucially, FDS adds more interdisciplinary, methodological variety and technological components to the analysis process, thus generating the need for a holistic view in the management of an FDS project. In this article, the author presents a holistic and structured approach to efficiently and effectively manage an FDS project. The main contribution of the article is to provide a comprehensive,

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

financial-domain–oriented approach to FDS management that can be used by researchers and practitioners to exploit the opportunities arising from the evolving finance landscape.

TOPICS: ***Statistical methods, simulations, big data/machine learning****

PCA for Implied Volatility Surfaces 85

Marco Avellaneda, Brian Healy, Andrew Papanicolaou, and George Papanicolaou

Principal component analysis (PCA) is a useful tool when trying to construct factor models from historical asset returns. For the implied volatilities of US equities, there is a PCA-based model with a principal eigenportfolio whose return time series lies close to that of an overarching market factor. The authors show that this market factor is the index resulting from the daily compounding of a weighted average of implied-volatility returns, with weights based on the options' open interest and Vega. The authors also analyze the singular vectors derived from the tensor structure of the implied volatilities of S&P 500 constituents and find evidence indicating that some type of open interest- and Vega-weighted index should be one of at least two significant factors in this market.

TOPICS: ***Statistical methods, simulations, big data/machine learning****

Evaluating Cost Efficiencies in Asset Management and Custody Banks Using Data Envelopment Analysis 110

D. K. Malhotra, Rashmi Malhotra, and Robert Nydick

As the aging baby boomer generation saves for retirement and demands suitable retirement savings vehicles, the role of asset and custody management banks is growing and the assets under their management are rising. The authors used a data envelopment analysis (DEA) model to evaluate the economies of scale in the asset management and custody bank subindustry of the capital market industry for the period of 2010 to 2017. They measured company size in two ways: total assets under management of the firm and total revenue generated by the firm. The authors evaluated cost efficiencies with cost measured in terms of total expenses, and they evaluated two important components of total expenses: total distribution expenses and general administrative and selling expenses. Using a returns to scale (RTS) DEA model, they found that most asset management firms initially exhibit increasing RTS until they become too large. The trend persisted when general and administrative expenses were evaluated and when total distribution expenses were evaluated.

TOPICS: ***Portfolio management/multi-asset allocation, portfolio construction, wealth management, retirement****

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

It's All About Data: *How to Make Good Decisions in a World Awash with Information*

MEHRZAD MAHDAVI AND HOSSEIN KAZEMI

MEHRZAD MAHDAVI is executive director of the FDP Institute in Amherst, MA.
mmahdavi@fdpinstitute.org

HOSSEIN KAZEMI is the Philipp Distinguished Professor of Finance in the Isenberg School of Management at the University of Massachusetts—Amherst and a senior advisor at the CAIA Association in Amherst, MA.
hkazemi@caia.org

KEY FINDINGS

- The rise of alternative data in finance is creating major opportunities in all areas of the financial industry, including risk management, portfolio construction, investment banking, and insurance.
- To build trusted outcomes in AI/ML initiatives, financial professionals' roles are critical. Given the many nuances in using big data, there is a need for vetted protocols and methods in selecting data sets and algorithms. Best practices and guidelines are effective in reducing the risks of using AI/ML, including overfitting, lack of interpretability, biased inputs, and unethical use of data.
- Given the major shortage of talent in AI/data science in finance, practical training of employees and continued education are keys to scale roll out to enable future of finance.

ABSTRACT: *The rise of big and alternative data has created significant new business opportunities in the financial sector. As we start on this journey of fast-moving technology disruption, financial professionals have a rare opportunity to balance the exponential growth of artificial intelligence (AI)/data science with ethics, bias, and privacy to create* trusted *data-driven decision making. In this article, the authors discuss the nuances of big data sets that are critical when one considers standards, processes, best practices, and modeling algorithms for the deployment of AI systems. In addition, this industry is widely guided by a fiduciary standard that puts the interests of the client above all else. It is therefore critical to have a thorough understanding of the limitations of our knowledge, because there are many* known unknowns *and* unknown unknowns *that can have a significant impact on outcomes. The authors emphasize key success factors for the deployment of AI initiatives: talent and bridging the skills gap. To achieve a lasting impact of big data initiatives, multidisciplinary teams with well-defined roles need to be established with continuing training and education. The prize is the finance of the future.*

TOPICS: ***Simulations, big data/machine learning****

Most agree that data is the new currency. The currency world, however, is a trusted system with significant globally established infrastructure, standards, ecosystems, and regulations to support transactions. On the other hand, data—our emerging currency—has yet to be well understood and lacks documented best practices.

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

The world is awash with data, and it is easy to assume we have all we need to make good decisions. However, misunderstanding the nuances of this digital asset can lead to problems. Working with the big data–driven world, the quotation made famous by Donald Rumsfeld is a good guidepost: "There are known knowns, known unknowns, and then there are unknown unknowns." For example, missing data can lead to conclusions and actions that are wrong, dangerous, unethical, or even disastrous. Lack of standard processes, procedures, and controls can lead to increased risks and mistrust in data/artificial intelligence (AI)–driven results.

We already see examples of a lack of trust in data-driven decisions. Patients refuse to accept diagnoses made by AI systems even if they are correct; consumers are leery of credit decisions enabled by AI, suspecting bias. This potential lack of trust is a major hurdle in the deployment of AI/data science in the financial and other business sectors. According to Cao et al. (2019), relatively few investment professionals are currently using AI/big data techniques in their investment processes. Most portfolio managers continue to rely on Excel and desktop market data tools; only 10% of portfolio manager respondents have used AI/machine learning techniques in the past 12 months, even though they see AI as the future of investment.

A learning curve exists not only for the professional implementation of AI but also for the resulting audience of customers. The digital tsunami is upon us. There is no silver bullet or easy answers on how to build the required trust in infrastructure. To start, a certified set of standards, processes, practices, and technologies must be established, rather than the fragmented one-off projects and pilots that are common today. In addition, high ethical principles and a fiduciary standard that puts clients' interests above all else should guide us above and beyond rules and regulations. Otherwise, we will risk being drowned by the tsunami of data.

The future of investment management is driven by AI/data science, as called out by numerous surveys and research work discussed by Bailey (2019). Navigating the road to reach investment professionals of the future requires an understanding of the nuances of big data.

In this article, we review and recommend practices for AI/big data implementation, including the big data market, data types, data models and algorithms, challenges in selecting the right data for the financial sector, and the required training and skill set for the roles that will be required. We will conclude with a brief on how big data will affect the financial sector and a call to action for building trusted data-driven investments.

BIG DATA IS BIG BUSINESS

The global market for big data, related technology, and analytics is expected to grow to more than $200 billion in 2020, according to Trent (2019). The financial industry, with a 15% spending share, is one of the important drivers of this growth. This spending includes acquiring datasets, building big data technology, and hiring appropriate talent.

Trent (2019) described in detail the explosive growth in big data and machine learning in finance and how this is likely to play out in the future. Findings include the following:

- The use of big data and machine learning is spreading beyond the realm of a handful of quantitative hedge funds to other asset managers. Although still encompassing a small percentage of total potential users, the growth curve is steep.
- Total spending on alternative data by mutual funds, hedge funds, pension funds, and others who buy securities (for their own or clients' accounts) is projected to jump from $232 million in 2016 to $1.7 billion in 2020, according to Watts (2019).
- A broader measure of spending, including outlays for data sources, data science, information technology (IT) infrastructure, data management, and systems development, is expected to rise to more than $7 billion by this year, according to Bailey (2019).
- The 72% of financial firms that have tried alternative data analysis have seen an enhancement of their profitability as a result, and over 20% of those who achieved gains report that they obtain an average of 20% of their alpha from alternative data operations (Bailey 2019).

Some of the principal providers of alternative data are Planet Labs, DigitalGlobe, Planet IQ, Orbital Insight, Descartes Labs, RS Metrics, and Spire.

As the market for alternative data expands, three kinds of data intermediaries are emerging: consultants, data aggregators, and IT firms offering technology solutions to big data clients. The roles of IT firms are increasing as they provide public, private, or hybrid cloud architecture that allows clients to onboard data

quickly. Examples of such firms are IBM (Big Insights), SAP (HANA), and Amazon (AWS cloud services).

HOW WILL BIG DATA AND MACHINE LEARNING CHANGE THE INVESTMENT LANDSCAPE?

The changes will likely be profound. As more investors adopt alternative datasets, the market will start reacting faster and will increasingly anticipate traditional or old data sources, such as quarterly corporate earnings or low-frequency macroeconomic data. This change gives an edge to quant managers and those willing to adapt to and learn about new datasets and methods. Eventually, old datasets will lose most predictive value, and new datasets that capture big data will increasingly become standardized. There will be an ongoing effort to uncover new higher-frequency datasets and refine or supplement old ones. Machine learning techniques will become a standard tool for quantitative investors and perhaps some fundamental investors, too. Systematic strategies such as risk premiums, trend followers, and equity long–short quants will increasingly adopt machine learning tools and methods.

Hedge fund managers are positioned to get much value from big data. It is commonly believed that buy-side and especially long-term investors have little use for big data. Although big data may provide the most value for short-term investors, long-term institutional investors can benefit from big data. More importantly, they may not have to compete with hedge fund managers to acquire alternative data sets that are most suited for short-term trading. Monk, Prins, and Rook (2019) argued that long-term investors should focus on defensive strategies, which are defined as those that help them manage risk and increase operational efficiency. By focusing on defensive strategies, long-term investors should be able to acquire alternative data sets at reasonable costs because they may not necessarily be in high demand by alternative asset managers. For example, alternative data sets may help these long-term investors detect long-term trends that take place owing to changes in tastes, demographics, and other secular factors. The increased demand for organic and healthier food is an example of such secular trends that would be of value to long-term investors. Big data and machine learning algorithms could be used to improve compliance at the firm level and could be used to perform due diligence on managers hired by long-term investors.

As pointed out by Lopez de Prado in a recent Bloomberg article (Asmundsson 2018), the majority of the hedge funds that have consistently beaten the market averages in recent years have used highly mathematical, data-intensive strategies. The same article reported that several traditional hedge funds have closed, whereas quantitative funds such as Renaissance Technologies and Two Sigma are still attracting new assets and clients.

BIG DATA TYPES AND SCOPE

A significant portion of analytics is dedicated to scoping, cleaning, and preparing data for analysis. If this important step is not taken, one will face the "garbage in, garbage out" problem.

The *big* in big data stands in for three prominent characteristics:

- **Volume:** The size of data collected and stored through, for example, records, transactions, tables, and files.
- **Velocity:** The speed with which data are sent or received. Data can be streamed or received in batch mode and real-time or near real-time.
- **Variety:** Data often come from a variety of sources and in various formats—be it structured (e.g., SQL tables or CSV files), semi-structured (e.g., JSON or HTML), or unstructured (e.g., a blog post or video message).

Big data/alternative data in investment management can be further classified based on how it was generated: produced by individuals (e.g., social media posts), generated through business processes (e.g., e-commerce or credit card transactions data), or generated by sensors (e.g., satellite imagery, radar, sensors installed in industrial facilities or on equipment, drones). Data generated by individuals are often in an unstructured textual format and commonly require natural language processing. Sensor-generated data can be produced both in structured and unstructured formats. Many business-generated datasets, such as credit card transactions and company exhaust data, have common legal and privacy considerations.

Other attributes of data (aside from the preceding classification) may be of more interest to investment professionals, such as mapping datasets to an asset class or investment style, alpha content, quality of data, technical

specifications, and so on. The attribute of a dataset is its relevance for the following criteria:

- **Asset classes:** Most big data are still focused on equities and commodities. There is relatively little alternative data on interest rates and currencies, making such data sets more valuable to select investors.
- **Investment styles:** Most data are sector and stock specific and relevant for equity long–short investors. A significant amount of data also is relevant for macro investors (e.g., consumer credit, Chinese economic activity, shipping data). Certain alternative datasets can be used to substitute for traditional metrics of market risk, and some signals are relevant only for high-frequency quant traders. Perhaps the most important data attribute is its potential.
- **Alpha content:** Alpha content has to be analyzed relative to the purchase price and implementation cost of deploying the dataset.
- **How well known a dataset is:** The more broadly a dataset is known, the less likely it is to lead to a standalone strategy with a strong Sharpe ratio. Well-known public datasets, such as financial ratios (e.g., price/earnings, price-to-book), likely have fairly low alpha content and are not viable as standalone strategies. (They may still be useful in a diversified risk premiums portfolio.) Most big datasets will be less well known, and new datasets emerge frequently.
- **Stage of processing of data when acquired:** Fundamental investors prefer processed signals and insights rather than a large amount of raw data. Most big/alternative datasets come in a semiprocessed format. Its alpha content and mapping to tradable instruments, however, need to be assessed, and the economics driving seasonality, outliers, and so on need to be considered. Finally, raw data are likely to be of little use for most investors.
- **Quality of data:** This is especially important for data scientists and quants. Missing data or outliers are an important consideration. If data have been backfilled, the method of imputing missing values must be clear. It must be specified if the missing data were missing at random or had patterns.
- **Frequency and latency:** Frequency of data can be intraday, daily, weekly, or even lower frequency. The latency of data can be due to collection, operational, or legal constraints.

Classification and attributes are part of a comprehensive taxonomy framework that needs to be established to label target data and choose among categories of machine learning algorithms.

BIG DATA AND MODELING ALGORIMTHS

Traditional econometric and statistical techniques may not be suitable tools for extracting maximum value from big data. Machine learning algorithms, along with deep learning models, have features capable of delivering maximum value from big data (see Kolanovic and Krishnamachari 2017). Although many of these techniques have their roots in traditional econometrics and statistical techniques (e.g., regression), they have been adjusted and updated to focus on prediction. This shift is at the core of most machine learning algorithms: to use big data to train a model with the explicit goal of predicting future outcomes based on a set of new inputs.

The accuracy and robustness of models are basically judged on their predictive power and less attention is paid to whether the estimates have some desirable properties when underlying assumptions of the model are not entirely satisfied. For example, although econometricians can spend an enormous amount of time on small sample properties of certain statistics, a data scientist will be most concerned with the overfitting and out-of-sample performance of the model. In fact, overfitting is a major concern because big data may allow building of models that perfectly explain the past but have very poor performance in the future.

Big data allows the data scientist to train increasingly complex models, leading to computer systems that can perform tasks that traditionally have required human intelligence. More precisely, the Financial Stability Board defines machine learning as "a method of designing a sequence of actions to solve a problem, known as algorithms, which optimize automatically through experience and with limited or no human intervention." These techniques can be used to find patterns in big data from increasingly diverse and innovative sources.

The field of machine learning has made tremendous progress in recent years mostly because of a decline in the cost of storage and computing power. The availability of big data enables data scientists to train algorithms with a much higher degree of complexity than what was imagined only a few years ago. There are several categories of machine learning algorithms. These categories vary according to the level of human intervention required in labeling the target data.

- **Supervised learning:** In this algorithm class, a set of training data that contains labels assigned by humans on some portion of the observations is used to train the algorithm. For example, the data may consist of the content of emails that have been labeled as spam or not spam. Another example is a set of transaction data labeled as being either fraudulent or not fraudulent. The algorithm will learn a general rule of classification that it will use to predict the labels for a new set of data (e.g., new emails or transactions). More broadly, regression and classification are examples of supervised learning techniques.
- **Unsupervised learning:** This category often uses clustering algorithms and refers to the case in which the data provided to the algorithm do not contain labels. The algorithm is given the task of detecting patterns in the data by identifying those observations that share a similar set of characteristics that are of interest to the user. For example, information about the past performance of a set of hedge funds may be used to identify hedge funds that will likely perform similarly during a time of rising inflation. The results can be used to create a diversified portfolio of hedge funds.
- **Reinforcement learning:** These algorithms are a hybrid of supervised and unsupervised learning. For these algorithms, an unlabeled set of data is given to the algorithm, which is used to determine a course of action. The algorithm receives feedback that helps it learn. The feedback could be built into the algorithm or could come from a human. For example, reinforcement learning algorithms are widely used in robotics, self-driving cars, and games. A reinforcement learning algorithm was used to defeat the world's best players in the game of Go.
- **Deep learning:** This form of machine learning uses algorithms that employ structures designed to emulate the structure and function of the brain. Deep learning algorithms, whose structures are called artificial neural networks, can be used for supervised, unsupervised, or reinforcement learning.

Once an algorithm has been trained, it is important to evaluate its performance before using it in the real world. The availability of big data enables us to train highly complex algorithms with thousands of parameters. Thanks to big data, these complex models are capable of explaining the past with a high degree of accuracy. However, this accuracy may come at the expense of overfitted models that will perform poorly out of sample or in the future. Therefore, before such models are used in the real world, they should be rigorously evaluated.

A number of methods have been developed to evaluate the performance of these algorithms. A full discussion is beyond the scope of this review. However, there are three widely used approaches:

- **Error analysis:** This approach is used to evaluate the performance of classification algorithms. It uses four evaluation metrics to evaluate a classification algorithm: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). These four metrics are typically used to create a two-by-two table known as a confusion matrix.
- **Receiver operating characteristic:** This approach provides a visual method for the evaluation of classification methods. The method involves a curve in which the relationship between (x-axis) the FP rate, defined as FP/(TN + FP) and the (y-axis) TP rate, defined as TP/(TP + FN) is examined.
- **Root mean squared error (RMSE):** The preceding methods are primarily used to measure the performance of regressions or algorithms that perform classification tasks. RMSE is defined as the sum of squared differences between the value predicted by the model and the actual value of the variable (e.g., predicted stock return versus actual stock return).

BIG DATA IMPLEMENTATION CHALLENGES

Implementing a big data strategy is not without challenges. A wide variety of concerns, which involve different teams of people, must be considered. Some of the major hurdles in scale deployment of data-driven decisions in the financial sector are as follows.

When building a big data strategy and putting a dataset into production, one should ensure that all legal and compliance requirements are satisfied. These requirements include potential issues around obtaining data and terms of use, potential material nonpublic

information (MNPI) in the dataset, and potential personally identifiable information in a dataset. Data license agreements and terms of use need to be fully understood (e.g., some data may be free for academic use but not for commercial use; data licenses may limit the number of users/teams allowed to access data; alternative data obtained through scraping of the web may in some cases face certain legal challenges).

Make sure that data do not contain MNPI. Most big data are in the public domain, and individual datasets, in most cases, do not contain material information. Only through sophisticated analysis of different datasets can one build an investment thesis (mosaic theory). For instance, geolocation foot traffic for a sample of retail stores that an analyst may use to model sales is free of MNPI. However, obtaining from all stores actual sales data that are otherwise not available publicly would be problematic. In the absence of an industrywide standard, refer to the US National Institute of Standards and Technology's guide to protecting the confidentiality of personally identifiable information (see McCallister, Grance, and Scarfone 2010). Signal decay is another challenge for big data because the information content (signals) may weaken in time owing to changes in transaction costs/capacity and so forth.

Costs of alternative datasets vary widely: Sentiment analysis can be obtained for a few hundred or thousand dollars, whereas comprehensive credit card data can cost up to a few million dollars a year. Trading strategies based on alternative data are tested, and alpha is estimated from a backtest. These tests can find whether a dataset has enough alpha to make it a viable standalone or portfolio-level trading strategy. Investors should not be surprised to come across alternative datasets with no alpha content.

Relevance and quality of the data are not straightforward, and there is little standardization for most data offerings. Initially, one should gather anecdotal intelligence on how well known and relevant the dataset is. Then, one should scrutinize the quality of the dataset (completeness, outliers, sampling methodology) to understand the level of preprocessing and various technical aspects of the dataset, which were discussed in the previous section. Finally, a trading strategy based on the dataset needs to be designed and tested. The backtest should be performed over different time periods and under different transaction cost assumptions. As with any quantitative strategy, special attention should be paid to avoid overfitting and in-sample biases. In these various steps (from acquiring data to trading implementation of a strategy), managers often partner with various participants in the big data industry.

AI system context is often too specific, and nuances must be determined. According to a recent blog by Mani (2019), an AI system today is not general-purpose enough to work facilely across multiple contexts and geographies. This issue can best be thought of as a triumvirate: data, domain knowledge, and techniques (or algorithms). Faces of local people, speech reflecting the regional language and dialect, and scenes with everyday endemic objects are what will be readily available as inputs to a retrieval or reasoning system. If any training is involved in teaching the system, it is best to use data similar to what the system will encounter when deployed.

Likewise, domain or subject matter expertise has many rich nuances—for example, the side of the street on which vehicles will be typically driven, the denomination and appearance of a banknote, or the uniform of a state's policeman. Algorithms can then aid in partially or completely automating driving, paying and checking out at a retail store, or directing people—via their smartphones—to a law enforcement officer for any clarifications and guidance in an emergency. Although advances in search and machine learning techniques have come from many countries via global teams of researchers and are widely published, often the systems have to be re-created domestically in the country of use, with local data and domain knowledge. Such local data and domain knowledge embedding aid system performance and robustness.

BIG DATA MASSIVE TRAINING AND NEW SKILL DEVELOPMENT

An estimated 80% of companies are already investing in AI, and most are facing challenges in hiring the talent they need to implement a useful AI application or product, according to Beck, Davenport, and Libert (2019). Clearly, there is an intensely competitive market for artificial intelligence and machine learning specialists. Many companies first attempt to hire PhD-level data scientists with expertise in AI algorithms and feature engineering. Some analysts have even equated AI talent with such researchers.

However, AI talent goes far beyond experts in machine learning. Equally important and less understood is the set of talent issues emerging around AI product development and engineering. Most firms have not filled these roles, and their AI projects are suffering as a result (see Henke et al. 2018).

The AI engineer role has already been realized for its importance for the practical deployment of AI systems. The key takeaway is that most companies need engineers to help develop products and production applications, rather than a researcher to help push the boundaries of AI technique and technology. These engineering skills include creating technology architectures that scale, writing and deploying bulletproof software, and integrating AI capabilities with existing systems. The people in AI engineering roles need to know something about AI but just as much about programming, computing, and corporate IT environments. Such skills are becoming increasingly important over time as AI knowledge and tools mature and algorithms and techniques become commoditized.

Another role, the AI "data czar" is becoming increasingly important. The machine learning race is increasingly a data race in which unique data, rather than cutting-edge modeling, is what creates a valuable AI solution. Unfortunately, sourcing and managing data is a skill set that does not often overlap with algorithm development. The AI data czar is typically a position that is created over time through experience, rather than hired out of school, although education in computer science or statistics can be very helpful. The role encompasses such capabilities as

- knowing what data sources are useful to address an AI question or problem
- being aware of how data are used in algorithms
- assessing data quality
- cleaning and treating data
- having a focus on detail (and being a stickler for data quality)
- possessing the strength to push back at technical teams
- knowing the typical ways to transform data

AI groups also need a role at the intersection of business strategy and AI methods, the AI translator. Such a person, usually a somewhat senior executive, can translate strategic objectives and business models into the types of AI that can advance them. Unfortunately, the role of a business leader with some understanding of AI techniques is rarely discussed and even less often filled.

Lack of multidisciplinary roles and expertise for AI development often results in either off-target or sunk-cost projects wherein the technology investment does not yield the return on investment anticipated by the board or the leadership team. According to Beck, Davenport, and Libert (2019), a solid AI product that provides customer, employee, operational, or investor value is about 40% problem and product definition; 40% data sourcing, cleaning, filling, and merging; and only 20% algorithm development.

Solving these problems requires an ongoing partnership between business and technology. Yet, most companies do not have a clear point of view about how AI can help organizations make better, more informed, and faster decisions or smarter products and services. Automating parts of decision-making and product development requires a person who can work at the intersection of strategy, business models, code development, algorithm creation, and product development—a rare breed.

Recruiting an AI quarterback to provide the business input, and ensuring success with well-defined metrics, is the most important job that most companies miss entirely. The businessperson who fills this role does not need to become an expert programmer, know the best AI tools from vendors, or delve into the nuances of neural networks versus logistic regression. He or she does, however, need to understand the basics of how different types of AI work and the data sets that will be deployed with them. Such individuals should also have a desire to get deeply involved and work iteratively with the AI team rather than throwing requirements over the wall, leaving the machine learning team with tough decisions. In addition, this role needs to create a clear economic use case and product road map that produces value for customers, employees, partners, or investors. In most cases, these individuals should lead the AI group, and the researchers, engineers, and data czar should report to them.

A CALL TO ACTION

In conclusion, it is important to recognize that big data, if it is to be truly useful in the finance

world, requires significantly more effort than merely downloading some large datasets from some public (or private) repository. A machine is only as intelligent as the data from which it learns. The more comprehensive the training data, the more generalized the machine will process new events, thereby mitigating common pitfalls such as overfitting and bias.

The transition to a big data framework will not be without setbacks. Certain types of data may lead to blind alleys—datasets that do not contain alpha and signals that have too little investment capacity, decay quickly, or are simply too expensive to purchase. Managers may invest too much into unnecessary infrastructure and, for example, build complex models and architecture that do not justify marginal performance improvements. Working with the big data-driven world, it is key to remember: "There are known knowns, known unknowns, and then there are unknown unknowns."

Understanding the nuances of big data datasets is critical and will consume most of the work required to produce AI-based analytics in the financial sector. Best practices in classification and attributions are required to produce reliable analytics for financial data. We believe an industrywide set of recommended practices and certifications, such as the one offered by the FDP Institute, is required to ensure trusted operations for big data–driven decisions. Certifications would be the guidepost for the development of a robust big data ecosystem involving specialized firms that collect, aggregate, and sell new datasets and research teams on both the buy-side and sell-side that evaluate data.

Finally, as discussed, a key factor in the successful deployment of AI initiatives is talent. In addition to the data scientist, critical talent will come from teams with multidisciplinary roles including the AI engineer, AI data czar, and AI translator working at the intersection of business strategy and data methods. These roles represent new AI/data science–related skills and jobs that are critical for the financial professional of the future. To achieve this goal, significant reskilling, training, and education is required. Many types of financial analysts, however, are in a good position to develop into these new roles—a win/win for both financial professionals and the employers. Reskilling and development of financial professionals ensure sustainable growth in the fastmoving field of *trusted* data-driven decision making.

REFERENCES

Asmundsson, J. "The Big Problem with Machine Learning Algorithms." *Bloomberg News*, October 18, 2018, https://www.bloomberg.com/news/articles/2018-10-09/the-big-problem-with-machine-learning-algorithms.

Bailey, D. H. "Go with Dig Data and Machine Learning, or Leave Finance to Those Who Do." *Mathematical Investor*, September 16, 2019, https://mathinvestor.org/2019/09/go-with-big-data-and-machine-learning-or-leave-finance-to-those-who-do/.

Beck, M., T. H. Davenport, and B. Libert. "The AI Roles Some Companies Forget to Fill." *Harvard Business Review*, March 14, 2019, https://hbr.org/2019/03/the-ai-roles-some-companies-forget-to-fill.

Cao, L., R. Fender, R. Stammers, and R. Urwin. "Investment Professional of the Future: Changing Roles, Skills, and Organizational Structures." Research and analysis, CFA Institute, 2019.

Henke, N., J. Levine, and P. McInerney. "You Don't Have to Be a Data Scientist to Fill This Must-Have Analytics Role." *Harvard Business Review*, February 5, 2018, https://hbr.org/2018/02/you-dont-have-to-be-a-data-scientist-to-fill-this-must-have-analytics-role?referral=03759&cm_vc=rr_item_page.bottom.

Kolanovic, M., and R. T. Krishnamachari. "Big Data and AI Strategies Machine Learning and Alternative Data Approach to Investing." Global Quantitative and Derivative Strategy, JPMorgan, 2017.

Mani, G. "What is made in India AI?" Rediff.com, November 23, 2019, https://www.rediff.com/business/column/what-is-made-in-india-ai/20191125.htm.

McCallister, E., T. Grance, and K. Scarfone. "Guide to Protecting the Confidentiality of Personally Identifiable Information." Report of the US Department of Commerce, National Institute of Standards and Technology, April 2010, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf.

Monk, A., M. Prins, and D. Rook. 2019. "Rethinking Alternative Data in Institutional Investment." *The Journal of Financial Data Science* 1 (1): 14–31.

Trent, N. "Global Big Data Market 2019 To Reach Valued At $200 billion And Grow At A 22.4% CAGR Forecast To 2024." Market Watch Press Release, April 2019, https://www.marketwatch.com/press-release/global-big-data-market-2019-to-reach-valued-at-200-billion-and-grow-at-a-224-cagr-forecast-to-2024-2019-04-05.

Watts, W. "The Explosion of Alternative Data Gives Regular Investors Access to Tools Previously Employed Only by Hedge Funds." Market Watch, September 19, 2019, https://www.marketwatch.com/story/the-explosion-of-alternative-data-gives-regular-investors-access-to-tools-previously-employed-only-by-hedge-funds-2019-09-05.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Rethinking Alternative Data in Institutional Investment

Ashby Monk, Marcel Prins, and Dane Rook

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/14

ABSTRACT: *As alternative data steadily become mainstream in finance, institutional investors may benefit from rethinking how they engage with alternative datasets. Specifically, they could gain from rethinking (1) alternative data's value proposition, (2) how they characterize alternative data, and (3) how they access alternative data. Rethinking their approaches to alternative data in these ways can help investors select alternative datasets that better align with their organizational resources and contexts. Such rethinking offers the greatest advantages when it focuses on building* defensive *and* defensible *strategies around alternative data, rather than prioritizing quicker exploitation of short-lived opportunities. Rethinking alternative data will require institutional investors to investigate new partnering possibilities, which should help them weather (and even thrive during) the escalating arms race among financial-market participants for alternative data. Building capacity for alternative data in these ways could also help investors accelerate innovation.*

Machine Learning in Asset Management—*Part 2: Portfolio Construction—Weight Optimization*

Derek Snow

Derek Snow
is a doctoral candidate of finance at the University of Auckland in Auckland, New Zealand.
d.snow@firmai.org

KEY FINDINGS

- Machine learning can help with most portfolio construction tasks, such as idea generation, alpha factor design, asset allocation, weight optimization, position sizing, and the testing of strategies.
- This is the second of a series of articles dealing with machine learning in asset management and, more narrowly, weight optimization strategies equipped with machine learning.
- Following from the previous article, different weight optimization methods are considered for supervised, unsupervised, and reinforcement learning frameworks.

ABSTRACT: *This is the second in a series of articles dealing with machine learning in asset management. This article focuses on portfolio weighting using machine learning. Following from the previous article (Snow 2020), which looked at trading strategies, this article identifies different weight optimization methods for supervised, unsupervised, and reinforcement learning frameworks. In total, seven submethods are summarized, with the code made available for further exploration.*

TOPICS: ***Big data/machine learning, analysis of individual factors/risk premia, portfolio construction, performance measurement****

In recent years, we have seen a trend toward online or dynamic portfolio asset allocation methods, clustering techniques, and methods that combine metadata on financial assets to help decide on the final weighting. Trading strategies and weight optimization methods can generally be considered as part of an integrated system. A trading strategy here is the use of perceived signals to execute trades, and weight optimization is the optimal weight allocation of active and passive strategies.

Strong parallels can be drawn between weight optimization in machine learning and weight optimization in portfolio management. When working with neural networks, we must often choose which optimization algorithm will produce faster and better updates for the network's weight and bias parameters. This is needed for the internal nodes to learn the optimal relationship in data to minimize a prespecified loss function. Researchers end up testing the performance of various optimizers, such as gradient descent, stochastic gradient descent (SGD), AdaGrad, RMSProp, and Adam, on their dataset. In the process, they might come to conclude that nothing really works that well. After some tinkering, a researcher might add a momentum component to SGD to curb the

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

high variance oscillation that makes standard SGD hard to converge. The next researcher might come along and see momentum as a problem when it ceases to stop as it reaches a local minimum. In response, the researcher could suggest a slope jump out of momentum as a reasonable correction.[1]

A development out of this constant tinkering was Adam, a method developed by Kingma and Ba (2014) that performs best by combining the properties of AdaGrad and RMSProp algorithms for a solution that can handle sparse gradients on noisy problems. In a similar spirit, we have seen modern portfolio optimization move from Markowitz's mean–variance (MV) portfolio using historical returns to dynamic models that use state-of-the-art reinforcement learning techniques. This process is not over; in both fields, new methods are brought to light on a frequent basis. Similar to machine learning, the optimal weighting strategy can and should be tested on validation data—in our parlance, back tests. We want to make sure that our data drive our optimization choices.

SUPERVISED LEARNING

Supervised learning techniques learn the relationship between independent attributes and a designated dependent attribute. Supervised learning refers to the mathematical structure describing how to make a prediction y_i given x_i. Instead of learning from the environment, as in reinforcement learning, supervised learning methods learn the relationships in data. All supervised learning tasks are divided into classification or regression tasks. Classification models are used to predict discrete responses (e.g., binary 1, 0; multiclass 1, 2, 3). Regression is used for predicting continuous responses (e.g., 3.5%, 35 times, $35,000). In portfolio weight optimization, we are generally presented with a continuous response problem.

Deep Portfolio

A useful deep learning technique for finance is autoencoders. This is a learning process to train the architecture to replicate X itself, namely $X = Y$, via a bottleneck structure. An autoencoder can create a more cost-effective representation of X. Autoencoding is an unsupervised learning method because there is no target to predict. However, the low-dimensional encoding is used as an input to a supervised learning method to calibrate and map the inputs to the desired target. The autoencoder demonstrates that, in deep learning, it is not necessary to model the variance–covariance matrix explicitly because the model is already in predictive form. This portfolio optimization process follows four steps: (1) autoencoding, (2) calibrating, (3) validating, and (4) verifying. This method has been popularized by Heaton, Polson, and Witte (2017).

1. **Autoencoding:** Find the market cap, denoted by $F_W^m(X)$, that solves the regularization problem

$$\min_W \left\| X - F_W^m(X) \right\|_2^2 \text{ subject to } \|W\| \le L^m$$

For appropriately chosen F_W^m, this autoencodes X with itself and creates a more information-efficient representation of X.

2. **Calibrating:** For a desired target Y, find the portfolio map, denoted by $F_W^p(X)$, that solves the regularization problem

$$\min_W \left\| Y - F_W^p(X) \right\|_2^2 \text{ subject to } \|W\| \le L^P$$

This creates a nonlinear portfolio from X for the approximation of objective Y.

3. **Validating:** Find L^m and L^P to suitably balance the trade-off between the two errors

$$\varepsilon_{m=} \left\| \hat{X} - F_{W_m}^m(\hat{X}) \right\|_2^2 \text{ and } \varepsilon_P = \left\| \hat{Y} - F_{W_P}^P(\hat{X}) \right\|_2^2$$

where W_m and W_p are the solutions to the validation and verification step

4. **Verifying:** Choose market cap F^m and portfolio map F^P such that the validation step is satisfactory.[2]

Linear Regression

We can also apply linear models to the problem of finding optimal portfolios (Britten-Jones 1999). There is a deep connection between fitting linear models and portfolio optimization. The normal equation is $\hat{\theta} = (X'X)^{-1}X'y$. Note that, in statistics or econometrics, it is more common to see β(beta) instead of θ(theta).

[1] Nesterov accelerated gradient.

[2] Resources: data (https://drive.google.com/open?id=1bJcUZbrZ8HFXs-cd0vGHeMop16Vf3n23); code (https://drive.google.com/open?id=1-hOEAiJqaNTUYIyamj26ZvHJNZq9XV09).

The normal equation makes sense if you do want a linear model and the number of explanatory variables (features) is not too big (e.g., < 100,000)—perfect for a small universe of stocks.

It is not necessary to use the normal equation; one can also solve for the best coefficients using numerical search such as gradient descent. With this portfolio optimization method, regularization becomes a critical tool to improve out-of-sample performance. The tested regularization methods include l2-norm (ridge regression), L1-norm (LASSO), and a combination of L1 and L2 norms (elastic net regression). Regularization effectively helps one to manage the variance–bias trade-off. Britten-Jones (1999) demonstrated a close equivalence between Markowitz portfolio optimization and ordinary least squares (OLS) regression.

Specifically, regress the excess returns of N stocks on a constant = 1 without an intercept. Collect the coefficients on each of the stocks in θ. Rescale θ so that the sum of the elements in θ is 1: ($\theta^* = \frac{\theta}{1^T\theta}$). The rescaled coefficients are the Markowitz optimal portfolio weights. As a result, any machine learning approach that yields coefficients for a linear model can be reinterpreted as a portfolio optimization approach (e.g., ridge regression → Tikhonov regularized optimal portfolio).[3]

Bayesian Sentiment

Advances in sentiment mining techniques led researchers in the 2010s to consider the use of public opinion for stock market prediction. Xing et al. (2018) decided to do the same for asset allocation by accounting for public mood using an online Bayesian asset allocation model. This model straddles between a trading strategy and weight optimization and falls within the supervised learning section because of its use of two neural models to generate market views.

To address the much-discussed limitations of the Markowitz model, Bayesian methods take additional information such as investor judgement and market fundamentals into account, a method once proposed by Black and Litterman (1990). The difference is that classical methods rely on financial experts and overlook public opinion and sentiment. A fair criticism of the Black and Litterman model is its subjective nature on market views; it leaves unanswered the question of how to assess these views. With the proposed method, public sentiment from the web can help researchers to formalize market views automatically.

The Black and Litterman model assumes that equilibrium returns are normally distributed $r_{eq} \sim N(\Pi, \tau\Sigma)$, where Σ is the covariance matrix of asset returns and τ is an indicator of the confidence level of the capital asset pricing model estimation of Π, the equilibrium risk premium of the market. The market views on the expected returns held by an investor agent are also normally distributed as $r_{views} \sim N(Q, \Omega)$. The posterior distribution of the portfolio returns that provide the views are therefore also Gaussian.

With the distributions denoted as $r_{BL} \sim N(\bar{\mu}, \bar{\Sigma})$, the vector of expected returns $\bar{\mu}$ and the covariance matrix $\bar{\Sigma}$ will be a function of the aforementioned variables:

$$[\bar{\mu}, \bar{\Sigma}] = f(\tau, \Sigma, \Omega, \Pi, Q)$$

The function can be induced by applying Bayes' theorem on the probability density function of the posterior expected returns:

$$pdf(\bar{\mu}) = \frac{pdf(\bar{\mu}|\Pi)\,pdf(\Pi)}{pdf(\Pi|\bar{\mu})}$$

The optimized Bayesian portfolio weights now have a form similar to the Markowitz model by substituting the MV version of Σ and μ with the new variables $\bar{\Sigma}$ and $\bar{\mu}$:

$$w^*_{BL} = (\delta\bar{\Sigma})^{-1}\bar{\mu}$$

A time series of asset prices, trading volume, and public sentiment data is used to approximate optimal market views. The sentiment is computed from a range of social media platforms using natural language processing techniques. The standard deviation will be interpreted as the confidence about the expected return of the portfolio, and a relative view would be described as taking the form "I have ω_1 confidence that asset x will outperform asset y by a%." An absolute view, in contrast, will take the form "I have a ω_2 confidence that asset z will outperform the market by b%." As a result, a portfolio of n assets and a set of k views can be represented by three matrixes $P_{k,n}$, $Q_{k,1}$, and $\Omega_{k,k}$.

[3] Resources: code (https://drive.google.com/open?id=1YDZQvz6Pn2AFDX2Uprfaq9JoGvk7RpJy); paper (Britten-Jones 1999).

Apart from performing slightly better than traditional methods, it also allows one to tell a nice Bayesian story.[4]

UNSUPERVISED LEARNING

Unlike supervised learning, which finds patterns using both input data and output data, unsupervised learning methods find patterns using only input data. An unsupervised learning framework is useful when researchers are not quite sure what to look for. Such a framework is often used for exploratory analysis and problem discovery purposes. Most unsupervised learning techniques take the form of dimensionality reduction or cluster analysis to establish groups of data that have some measure of similarity based on characteristic values.

Two of the most important techniques are K-means clustering and principal component analysis (PCA). The only requirement to be called an unsupervised learning strategy is to learn a new feature space that captures the characteristics of the original space by maximizing some objective function.[5] PCA attempts to reduce the number of features while preserving the variance, whereas clustering reduces the number of data points by summarizing them according to their mean expectations. However, those cluster assignments can also be used to label each data point with its assigned cluster, leading to a dimensionality reduction toward only one feature. K-mean and PCA, in some sense, maximize a similar objective function, with K-means having an additional categorical constraint.

PCA

PCA and clustering techniques are used to build classes of similar assets. The steps here involve integrated portfolio selection and risk estimation to optimize the portfolio. Let $Y = (Y_1, \ldots, Y_n)^T$ denote an n-dimensional random vector with variance–covariance matrix Σ. The goal of PCA is to construct linear combinations $P_i = \Sigma^n_{j=1} W_{ij} Y_j$, for $i = 1, \ldots n$ in such a way that the P_i values are orthogonal so that $\Sigma[P_i P_j] = 0$ for $i \neq j$ so that the P_i are ordered to explain the largest percentage of the total variability in the system.

Each P_i explains the most significant percentage of total variability in the system that has not already been explained by $P_i, \ldots, P_i - 1$. Using various selection techniques, one can identify an optimum level of these components and apply it to a universe of stocks. This is not a complicated method, but it is very powerful and can be used as a mediating step in various trading strategies, if not in the final allocation decision making. In the code notebook, there is additional experimentation to compare this technique with hierarchical clustering methods.[6]

HIERARCHICAL RISK PARITY

The problem, as stated by López de Prado (2016), is that the MV portfolios are optimal in sample (training set) but perform poorly out of sample (test set). One way to deal with this problem is to drop forecasts altogether (e.g., risk parity [RP]). The problem is that both RP and MV require the inversion of a positive-definite covariance matrix. A new method is therefore suggested to overcome the matrix inversion and forecast issue, called hierarchical risk parity.

Hierarchical risk parity works by grouping similar investments into clusters based on a distance metric. The covariance matrix's rows and columns also are reorganized so that the largest values lie along the diagonal. Finally, the allocations are split through recursive bisections of the reordered covariance matrix.

One starts by defining a distance measure between investments from zero and one, $d_{i,j}$, after which one clusters the pair of columns (i^*, j^*) together such that $(i^*, j^*) \text{argmin}_{(i,j)} = \{\grave{d}_{i,j}\}$ and $i \neq j$. The next step is to update $\{\grave{d}_{i,j}\}$ with the new cluster and apply steps 3 and 4 recursively until all $N - 1$ clusters are formed.

We now place correlated investments close together and uncorrelated investments further apart and carry out a top-down allocation: We assign unit weights to all items by recursively bisecting a list of items by computing the variance and the split factor and rescaling the allocations; we iterate this process until full allocation is achieved. In the code, I have included an implementation using Robert Martin's PyPortfolioOpt[7] and an implementation of Chapter 16 in *Advances in*

[4] Resource: code (https://colab.research.google.com/drive/1sMAoJZuuNIRnrivAzxHV5fulMOWO17mb).

[5] Inherent to the PCA is the maximization of variance through a simple linear algebra operation by taking the eigenvectors of a covariance matrix of features.

[6] Resource: code (https://colab.research.google.com/drive/1mm9r6EZOERHYkycDbc74GY7S2U6h1oTc).

[7] https://github.com/robertmartin8/PyPortfolioOpt.

Financial Machine Learning (López de Prado 2018) by Hudson & Thames,[8] a buy-side open-source research unit headed by Jacques Joubert.[9]

Network Graph

Given N assets in a portfolio, we have to identify the weights $w_i, (\Sigma_{i=1}^{N} w_i = 1)$ so that highly correlated assets obtain lower relative weights. To do this, we can use a weighted graph that is an ordered tuple, $G = (V, E, W)$, where V is a set of vertices (or nodes), E is a set of pairwise relationships (the edges) between the vertices, and W is a set of numerical values assigned to each edge. A useful representation of G is the adjacency matrix:

$$A_{ij} = \begin{cases} 1, & \textit{if } i \textit{ is adjacent to } j \\ 0, & \textit{otherwise} \end{cases}$$

Here the pairwise relations are expressed as the *ij* entries of an $N \times N$ matrix where N is the number of nodes. The strategy is then to transform the historical pricing data into a graph with edges weighted by the correlation between each stock. We can then use graph centrality measures and graph algorithms to obtain the desired allocation weights. In five steps, we do the following:

1. Compute the distance correlation matrix $\rho_D(X_i, X_j)$ for the open, high, low, close, and return time series.
2. Use the NetworkX[10] module to transform each distance correlation matrix into a weighted graph.
3. Adopt a winner-takes-all method and remove edges with correlations below a threshold value of $\rho_c = 0.325$ (adjust this threshold value if the graph disconnects):

$$Cor_{ij} = \begin{cases} \rho_D(X_i, X_j), & \rho \geq \rho_c \\ 0, & \textit{otherwise} \end{cases}$$

4. Inspect the distribution of edges (the so-called *degree distribution*) for each network.[11] The degree of the *i*th vertex is given as

$$Deg(i) = \sum_{j=1}^{N} A_{ij}$$

5. Finally, build a master network by averaging over the edge weights of the open, high, low, close, and return networks and derive the asset weights from its structure.

Multiple variations are available for constructing the network and obtaining the distance; a few different approaches are considered in the notebook. An extensive analysis can be found in a notebook by Maya Benowitz, a quant at CarVal.[12]

REINFORCEMENT LEARNING

Reinforcement learning in finance comprises the use of an agent that learns how to take actions in an environment to maximize some notion of cumulative reward. We have an agent that exists in a predefined environment. The agent receives as input the current state S_t and is asked to take an action A_t to receive a reward R_{t+1}, the information of which can be used to identify the next optimal action, A_{t+1}, given the new state S_{t+1}. The final objective function can be the realized/unrealized profit and loss and even risk-adjusted performance measures such as the Sharpe ratio. Allocation decisions in finance are challenging to deal with because finance is partially observed, nonstationary, regime dependent, and noisy. Standard models apply and recombine single-period predictions using an optimizer, but in the real world, the actions could have long-term consequences that could be acted against by the environment. Reinforcement learning can help us to deal with some of these problems by taking a more holistic approach.

Deep Determinist Policy Gradient

Deep learning models can be used in reinforcement learning to solve high-dimensional problems.

[8] https://hudsonthames.org/.

[9] Resources: data (https://drive.google.com/open?id=198fpHhD973i3rKa9D7oz-SrmBwPykQEc), code (https://drive.google.com/open?id=1z3Fe7QXZ6c566KOG3HtQEfCc84UAGwFf), code 2 (https://colab.research.google.com/drive/1-Z3OjjnIR-41E2tycKFosvxEt-RrAgZB).

[10] https://github.com/networkx/networkx.

[11] The degree of a node is simply the number of connections it has to other nodes.

[12] Code available at https://colab.research.google.com/drive/10WNiVuICvFajW2uTDrwI6w7aSUkjINPl.

To deal with the continuous action space of portfolio weighting, we can use Google DeepMind's off-policy and model free algorithm called deep deterministic policy gradient (Lillicrap et al. 2015). Using this method, we do not have to discretize continuous action spaces, which might lead to the curse of dimensionality and a loss of valuable information. The environment is derived from an OpenAI environment class.[13] This environment takes arguments like trading cost and window size into account to approach a realistic setting.

To start formalizing the problem, we set the number of stocks to N. At time step 1, we are fully invested, and the close/open relative price vectors are defined as

$$\gamma_t = \left[1, \frac{S_{1,t,close}}{S_{1,t,open}}, \frac{S_{2,t,close}}{S_{2,t,open}}, \ldots, \frac{S_{N,t,close}}{S_{N,t,open}}\right]$$

The portfolio weight vector is then defined as

$$w_t = [w_{o,t}, w_{1,t}, \ldots, w_{N,t}]$$

Here $w_{i,t}$ is the fraction of investment in stock i at time stamp t. Note that $w_{o,t}$ represents the fraction of cash maintained. The profit after timestamp T is

$$p_T = \prod_{t=1}^{T} \gamma_t \cdot w_{t-1}$$

Once we consider the trading cost factor of μ, the trading cost at each time stamp is

$$\mu_t = \mu \sum \left| \frac{\gamma_t \odot w_{t-1}}{\gamma_t \cdot w_{t-1}} - w_t \right|$$

where $\odot$ is the element-wise product. The profit can then be formulated as

$$p_T = \prod_{t=1}^{T} (1 - \mu_t) \gamma_t \cdot w_{t-1}$$

Now that we have dealt with the problem formulation, we have to formulate the Markov decision process. We set the state o_t as the fixed window W price history $\overrightarrow{S_{i,t}}$ of all the assets N:

$$o_t = [\overrightarrow{S_{1,t}}, \overrightarrow{S_{2,t}}, \ldots, \overrightarrow{S_{N,t}}]$$

$$\overrightarrow{S_{i,t}} = \begin{bmatrix} S_{i,t-W} \\ S_{i,t-W+1} \\ \vdots \\ S_{i,t-W} - 1 \end{bmatrix}$$

Moreover, the action a_t is merely the portfolio weight vector as previously defined, w_t. We therefore want to train a policy network: $\pi_\theta(a_t, o_t)$. The underlying evolution and transition of state are then determined by the market, and we can simply obtain the observation of the states. The reward is specified as the log-profit at each time step, $\log p_T$, which avoids the sparsity of rewards problem. Recall that with the deep deterministic policy gradient algorithm we want to learn the policy network $\pi_\theta(a_t, o_t)$ with a continuous action space reinforcement algorithm. The actor network is set to be the same as the policy network; the critic network is a linear combination of the actor network structure of the state and action; the exploration noise is set by the Ornstein–Uhlenbeck process with zero mean, 0.3 sigma, and 0.15 theta.[14]

The specific implementation in the code notebook looks at 15 stocks for the first 10 months in 2018 with data in minute format with open, close, high, low, and volume variables. This method is adapted from Jiang, Xu, and Liang's (2017) work published on arXiv. The action space contains a cash position, long positions, and short positions. The algorithm is hardcoded to only act every seven minutes. Note that reinforcement models can be very unstable and are generally hard to converge; when they do converge, they generally overfit. At this stage, without any further innovation, reinforcement learning models have to be implemented with great care.[15]

SUMMARY

There are several ways in which modern machine learning innovations can be used to help portfolio

[13] More can be found in their documentation: https://gym.openai.com/docs/.

[14] For more information on this distinction, see the original DeepMind paper referenced at the start.

[15] Resource: code (https://colab.research.google.com/drive/1L3-D2ZmGZkPRsB9gb5BviGkSkMTLti7_).

managers optimally allocate assets. Machine learning is poised to partially replace traditionally rigid allocation methods. This article divides these methods into supervised, unsupervised, and reinforcement learning frameworks. Within the supervised learning framework, three techniques were considered: first, a more traditional linear approach using OLS, Ridge, and LASSO regressions; second, a nonlinear deep learning approach using autoencoders; and, finally, a Bayesian sentiment method. The article further identified three unsupervised methods using principal component analysis, hierarchical clustering analysis, and network graphs. The article finished with a deep reinforcement learning approach to portfolio weight optimization. As with the first article, I expect to see much more growth in reinforcement learning methods in the coming years. To paraphrase Vladimir Vapkin, the inventor of the support-vector machine method: One should avoid solving difficult intermediate problems when what one truly wants to solve is a target problem.

REFERENCES

Black, F., and R. Litterman. 1990. "Asset Allocation: Combining Investor Views with Market Equilibrium." *Goldman Sachs Fixed Income Research* 115.

Britten-Jones, M. 1999. "The Sampling Error in Estimates of Mean–Variance Efficient Portfolio Weights." *The Journal of Finance* 54 (2): 655–671.

Heaton, J. B., N. G. Polson, and Witte. 2017. "Deep Learning for Finance: Deep Portfolios." *Applied Stochastic Models in Business and Industry* 33 (1): 3–12.

Jiang, Z., D. Xu, D., and J. Liang. 2017. "A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem." arXiv preprint 1706.10059.

Kingma, D. P., and J. Ba. 2014. "Adam: A Method for Stochastic Optimization." arXiv preprint 1412.6980.

Lillicrap, T. P., J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. 2015. "Continuous Control with Deep Reinforcement Learning." arXiv preprint 1509.02971.

López de Prado, M. 2016. "Building Diversified Portfolios That Outperform Out of Sample." *The Journal of Portfolio Management* 42 (4): 59–69.

——. *Advances in Financial Machine Learning*, 1st ed.. Newark: John Wiley & Sons, 2018.

Snow, D. 2020. "Machine Learning in Asset Management—Part 1: Portfolio Construction—Trading Strategies." *The Journal of Financial Data Science* 2 (1): 10–23.

Xing, F. Z., E. Cambria, L. Malandri, and C. Vercellis. 2018. "Discovering Bayesian Market Views for Intelligent Asset Allocation." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 120–135. Cham, Switzerland: Springer, 2018.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Machine Learning in Asset Management—*Part 1: Portfolio Construction—Trading Strategies*

Derek Snow
The Journal of Financial Data Science
https://jfds.pm-research.com/content/2/1/10

ABSTRACT: *This is the first in a series of articles dealing with machine learning in asset management. Asset management can be broken into the following tasks: (1) portfolio construction, (2) risk management, (3) capital management, (4) infrastructure and deployment, and (5) sales and marketing. This article focuses on portfolio construction using machine learning. Historically, algorithmic trading could be more narrowly defined as the automation of sell-side trade execution, but since the introduction of more advanced algorithms, the definition has grown to include idea generation, alpha factor design, asset allocation, position sizing, and the testing of strategies. Machine learning, from the vantage of a decision-making tool, can help in all these areas.*

Building Diversified Portfolios that Outperform Out of Sample

Marcos López de Prado
The Journal of Portfolio Management
https://jpm.pm-research.com/content/42/4/59

ABSTRACT: *In this article, the author introduces the Hierarchical Risk Parity (HRP) approach to address three major concerns of quadratic optimizers, in general, and Markowitz's critical line algorithm (CLA), in particular: instability, concentration, and underperformance. HRP applies modern mathematics (graph theory and*

machine-learning techniques) to build a diversified portfolio based on the information contained in the covariance matrix. However, unlike quadratic optimizers, HRP does not require the invertibility of the covariance matrix. In fact, HRP can compute a portfolio on an ill-degenerated or even a singular covariance matrix—an impossible feat for quadratic optimizers. Monte Carlo experiments show that HRP delivers lower out-of-sample variance than CLA, even though minimum variance is CLA's optimization objective. HRP also produces less risky portfolios out of sample compared to traditional risk parity methods.

Deep Reinforcement Learning for Trading

ZIHAO ZHANG, STEFAN ZOHREN, AND STEPHEN ROBERTS

ZIHAO ZHANG
is a D.Phil. student with the Oxford-Man Institute of Quantitative Finance and the Machine Learning Research Group at the University of Oxford in Oxford, UK.
zihao.zhang@worc.ox.ac.uk

STEFAN ZOHREN
is an associate professor (research) with the Oxford-Man Institute of Quantitative Finance and the Machine Learning Research Group at the University of Oxford in Oxford, UK.
zohren@robots.ox.ac.uk

STEPHEN ROBERTS
is the director of the Oxford-Man Institute of Quantitative Finance, the founding director of the Oxford Centre for Doctoral Training in Autonomous Intelligent Machines and Systems, and the Royal Academy of Engineering/Man Group Professor in the Machine Learning Research Group at the University of Oxford in Oxford, UK.
sjrob@robots.ox.ac.uk

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

KEY FINDINGS

- In this article, the authors introduce reinforcement learning algorithms to design trading strategies for futures contracts. They investigate both discrete and continuous action spaces and improve reward functions by using volatility scaling to scale trade positions based on market volatility.
- The authors discuss the connection between modern portfolio theory and the reinforcement learning reward hypothesis and show that they are equivalent if a linear utility function is used.
- The authors back test their methods on 50 very liquid futures contracts from 2011 to 2019, and their algorithms deliver positive profits despite heavy transaction costs.

ABSTRACT: *In this article, the authors adopt deep reinforcement learning algorithms to design trading strategies for continuous futures contracts. Both discrete and continuous action spaces are considered, and volatility scaling is incorporated to create reward functions that scale trade positions based on market volatility. They test their algorithms on 50 very liquid futures contracts from 2011 to 2019 and investigate how performance varies across different asset classes, including commodities, equity indexes, fixed income, and foreign exchange markets. They compare their algorithms against classical time-series momentum strategies and show that their method outperforms such baseline models, delivering positive profits despite heavy transaction costs. The experiments show that the proposed algorithms can follow large market trends without changing positions and can also scale down, or hold, through consolidation periods.*

TOPICS: ***Futures and forward contracts, exchanges/markets/clearinghouses, statistical methods, simulations****

Financial trading has been a widely researched topic, and a variety of methods have been proposed to trade markets over the last few decades. These include fundamental analysis (Graham and Dodd 1934), technical analysis (Murphy 1999), and algorithmic trading (Chan 2009). Indeed, many practitioners use a hybrid of these techniques to make trades (Schwager 2017). Algorithmic trading has arguably gained most recent interest and accounts for about 75% of trading volume in the US stock exchanges (Chan 2009). The advantages of algorithmic trading are widespread, ranging from strong computing foundations to faster execution and risk diversification. One key

component of such trading systems is a predictive signal that can lead to alpha (excess return); to this end, mathematical and statistical methods are widely applied. However, because of the low signal-to-noise ratio of financial data and the dynamic nature of markets, the design of these methods is nontrivial, and the effectiveness of commonly derived signals varies through time.

In recent years, machine learning algorithms have gained much popularity in many areas, with notable successes in diverse application domains, including image classification (Krizhevsky, Sutskever, and Hinton 2012) and natural language processing (Collobert et al. 2011). Similar techniques have also been applied to financial markets in an attempt to find higher alpha (for a few examples using such techniques in the context of high-frequency data, see Tsantekidis et al. 2017a; Zhang, Zohren, and Roberts 2018, 2019a, 2019b; Sirignano and Cont 2019). Most research focuses on regression and classification pipelines in which excess returns, or market movements, are predicted over some (fixed) horizons. However, there is little discussion related to transforming these predictive signals into actual trade positions (see Lim, Zohren, and Roberts 2019 for an attempt to train a deep learning model to learn positions directly). Indeed, such a mapping is nontrivial. As an example, predictive horizons are often relatively short (one day or a few days ahead if using daily data); however, large trends can persist for weeks or months with some moderate consolidation periods. We therefore need a signal that not only has good predictive power but also can consistently produce good directional calls.

In this article, we report on reinforcement learning (RL) (Sutton and Barto 1998) algorithms to tackle the aforementioned problems. Instead of making predictions—followed by a trade decision based on predictions—we train our models to directly output trade positions, bypassing the explicit forecasting step. Modern portfolio theory (Arrow 1971; Pratt 1978; Ingersoll 1987) implies that, given a finite time horizon, an investor chooses actions to maximize some expected utility of final wealth:

$$\mathbb{E}[U(W_T)] = \mathbb{E}\left[U\left(W_0 + \sum_{t=1}^{T} \delta W_t\right)\right] \tag{1}$$

where U is the utility function, W_T is the final wealth over a finite horizon T, and δW_t represents the change in wealth. The performance of the final wealth measure depends upon sequences of interdependent actions in which optimal trading decisions do not just decide immediate trade returns but also affect subsequent future returns. As mentioned by Merton (1969) and Ritter (2017), this falls under the framework of optimal control theory (Kirk 2012) and forms a classical sequential decision-making process. If the investor is risk neutral, the utility function becomes linear, and we only need to maximize the expected cumulative trades returns, $\mathbb{E}(\Sigma_{t=1}^{T}\delta W_t)$; we observe that the problem fits exactly with the framework of RL, the goal of which is to maximize some expected cumulative rewards via an agent interacting with an uncertain environment. Under the RL framework, we can directly map different market situations to trade positions and conveniently include market frictions, such as commissions, in our reward functions, allowing trading performance to be directly optimized.

In this article, we adopt state-of-the-art RL algorithms to the aforementioned problem setting, including deep Q-learning networks (DQN) (Mnih et al. 2013; Van Hasselt, Guez, and Silver 2016), policy gradients (PG) (Williams 1992), and advantage actor–critic (A2C) (Mnih et al. 2016). Both discrete and continuous action spaces are explored in our work, and we improve reward functions with volatility scaling (Moskowitz, Ooi, and Pedersen 2012; Harvey et al. 2018) to scale up trade positions when volatility is low, and vice versa. We show the robustness and consistency of our algorithms by testing them on 50 very liquid futures contracts (CLC Database 2019) between 2011 and 2019. Our dataset consists of different asset classes, including commodities, equity indexes, fixed income, and foreign exchange (FX) markets. We compare our method with classical time-series momentum strategies and find that our method outperforms baseline models and generates positive returns in all sectors despite heavy transaction costs. Time-series strategies work well in trending markets, such as fixed-income markets, but suffer losses in FX markets in which directional moves are less usual. Our experiments show that our algorithms can monetize on large moves without changing positions and deal with markets that are more mean reverting.

The remainder of the article is structured as follows. We first introduce the current literature and present our methodology. We then compare our methods with baseline algorithms and show how trading performance

varies among asset classes; a description of the dataset and training methodology is also included. At the end, we offer our conclusions and discuss extensions of our work.

LITERATURE REVIEW

In this section, we review some classical trading strategies and discuss how RL has been applied to this field. Fundamental analysis aims to measure the intrinsic value of a security by examining economic data so investors can compare the security's current price with estimates to determine whether the security is undervalued or overvalued. One of the established strategies is called CAN-SLIM (O'Neil 2009), and it is based on a major study of market winners from 1880 to 2009. A common criticism of fundamental analysis is that the entrance and exit timing of trades is not specified. Even when markets move toward the estimated price, bad timing in entering trades could lead to huge drawdowns, and such moves in account values are often not bearable to investors, shaking them out of the markets. Technical analysis is in contrast to fundamental analysis, in which a security's historical price data are used to study price patterns. Technicians place trades based on a combination of indicators such as the relative strength index (RSI) and Bollinger bands. However, owing to the lack of analysis on economic or market conditions, the predictability of these signals is not strong, often leading to false breakouts.

Algorithmic trading is a more systematic approach that involves mathematical modeling and automated execution. Examples include trend following (Schwager 2017), mean-reversion (Chan 2013), statistical arbitrage (Chan 2009), and delta-neutral trading strategies (Michaud 1989). We mainly review time-series momentum strategies by Moskowitz, Ooi, and Pedersen (2012) because we benchmark our models against their algorithms. Their work developed a very robust trading strategy by simply taking the sign of returns over the last year as a signal; they demonstrated profitability for every contract considered within 58 liquid instruments over 25 years. Thenceforth, a number of methods (Baz et al. 2015; Baltas and Kosowski 2017; Rohrbach, Suremann, and Osterrieder 2017; Lim, Zohren, and Roberts 2019) have been proposed to enhance this strategy by estimating the trend and mapping to actual trade positions. However, these strategies are designed to profit from large directional moves but can suffer huge losses if markets move sideways as the predictability of these signals deteriorates and excess turnover erodes profitability. Here, we adopt time-series momentum features along with technical indicators to represent state space and obtain trade positions directly using RL. The idea of our representation is simple: extracting information across different momentum features and outputting positions based on the aggregated information.

The current literature on RL in trading can be categorized into three main methods: critic-only, actor-only, and actor–critic approaches (Fischer 2018). The critic approach, mainly DQN, features most often in publications in this field (Tan, Quek, and Cheng 2011; Bertoluzzo and Corazza 2012; Jin and El-Saawy 2016; Ritter 2017; Huang 2018). A state-action value function, Q, is constructed to represent how good a particular action is in a state. Discrete action spaces are adopted in these works, and an agent is trained to fully go long or short a position. However, a fully invested position is risky during high-volatility periods, exposing one to severe risk when opposite moves occur. Ideally, one would like to scale positions up or down according to current market conditions. Doing this requires one to have large action spaces; however, the critic approach suffers from large action spaces because a score must be assigned for each possible action.

The second most common approach is the actor-only approach (Moody et al. 1998; Moody and Saffell 2001; Deng et al. 2016; Lim, Zohren, and Roberts 2019), in which the agent directly optimizes the objective function without computing the expected outcomes of each action in a state. Because a policy is directly learned, actor-only approaches can be generalized to continuous action spaces. In the works of Moody and Saffell (2001) and Lim, Zohren, and Roberts (2019), offline batch gradient ascent methods can be used to optimize the objective function, such as profits or Sharpe ratio, because the approach is differentiable end to end. However, this is different from standard RL actor-only approaches in which a distribution needs to be learned for the policy. To study the distribution of a policy, the policy gradient theorem (Sutton and Barto 1998) and Monte Carlo methods (Metropolis and Ulam 1949) are adopted in the training, and models are updated until the end of each episode. We often experience slow learning and need many samples to obtain an optimal policy because individual bad actions will be considered

good as long as the total rewards are good; thus, a long time is needed to adjust these actions.

The actor–critic approach forms the third category and aims to solve the aforementioned learning problems by updating the policy in real time. The key idea is to alternatively update two models in which one (the actor) controls how an agent performs given the current state, and the other (the critic) measures how good the chosen action is. However, this approach is the least studied method in financial applications (Li et al. 2007; Bekiros 2010; Xiong et al. 2018). We aim to supplement the literature and study the effectiveness of this approach in trading. For a more detailed discussion on state, action spaces, and reward functions, interested readers are pointed to the survey (Fischer 2018). Other important financial applications such as portfolio optimization and trade execution are also included in this work.

METHODOLOGY

In this section, we introduce our setups, including state, action spaces, and reward functions. We describe three RL algorithms used in our work, DQN, PG, and A2C methods.

Markov Decision Process Formalization

We can formulate the trading problem as a Markov decision process in which an agent interacts with the environment at discrete time steps. At each time step t, the agent receives some representations of the environment denoted as a state S_t. Given this state, an agent chooses an action A_t, and based on this action, a numerical reward R_{t+1} is given to the agent at the next time step, and the agent finds itself in a new state S_{t+1}. The interaction between the agent and the environment produces a trajectory $\tau = [S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots]$. At any time step t, the goal of RL is to maximize the expected return (essentially the expected discounted cumulative rewards) denoted as G_t at time t:

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \tag{2}$$

where γ is the discounting factor. If the utility function in Equation 1 has a linear form and we use R_t to represent trade returns, we can see that optimizing $\mathbb{E}(G)$ is equivalent to optimizing our expected wealth.

State space. In the literature, many different features have been used to represent state spaces. Among these features, a security's past price is always included, and the related technical indicators are often used (Fischer 2018). In our work, we take past price, returns (r_t) over different horizons, and technical indicators including moving average convergence divergence (MACD) (Baz et al. 2015) and the RSI (Wilder 1978) to represent states. At a given time step, we take the past 60 observations of each feature to form a single state. The following is a list of our features:

- Normalized close price series
- Returns over the past month, two months, three months, and one year are used. Following Lim, Zohren, and Roberts (2019), we normalize them by daily volatility adjusted to a reasonable time scale. As an example, we normalize annual returns as $r_{t-252,t}/(\sigma_t\sqrt{252})$, where σ_t is computed using an exponentially weighted moving standard deviation of r_t with a 60-day span,
- MACD indicators are proposed by Baz et al. (2015) where

$$\begin{aligned} \mathrm{MACD}_t &= \frac{q_t}{\mathrm{std}(q_{t-252:t})} \\ q_t &= (m(S) - m(L))/\mathrm{std}(p_{t-63:t}) \end{aligned} \tag{3}$$

  where $\mathrm{std}(p_{t-63:t})$ is the 63-day rolling standard deviation of prices p_t, and $m(S)$ is the exponentially weighted moving average of prices with a time scale S.
- The RSI is an oscillating indicator moving between 0 and 100. It indicates the oversold (a reading below 20) or overbought (above 80) condition of an asset by measuring the magnitude of recent price changes. We include this indicator with a look-back window of 30 days in our state representations.

Action space. We study both discrete and continuous action spaces. For discrete action spaces, a simple action set of {−1, 0, 1} is used, and each value represents the position directly (i.e., −1 corresponds to a maximally short position, 0 to no holdings, and 1 to a maximally long position). This representation of action space is also known as target orders (Huang 2018), in which a trade position is the output instead of the trading decision.

Note that if the current action and next action are the same, no transaction cost will occur and we just maintain our positions. If we move from a fully long position to a short position, transaction cost will be doubled. The design of continuous action spaces is very similar to the discrete case in which we still output trade positions but allow actions to be any value between −1 and 1 ($A_t \in [-1, 1]$).

Reward function. The design of the reward function depends on the utility function in Equation 1. Here, we let the utility function be profits representing a risk-insensitive trader, and the reward R_t at time t is

$$R_t = \mu\left[A_{t-1}\frac{\sigma_{tgt}}{\sigma_{t-1}}r_t - \text{bp}\ p_{t-1}\left|\frac{\sigma_{tgt}}{\sigma_{t-1}}A_{t-1} - \frac{\sigma_{tgt}}{\sigma_{t-2}}A_{t-2}\right|\right] \quad (4)$$

where σ_{tgt} is the volatility target and σ_{t-1} is an ex ante volatility estimate calculated using an exponentially weighted moving standard deviation with a 60-day window on r_t. This expression forms the volatility scaling of Moskowitz, Ooi, and Pedersen (2012), Harvey et al. (2018), and Lim, Zohren, and Roberts (2019), in which our position is scaled up when market volatility is low and vice versa. In addition, given a volatility target, our reward R_t is mostly driven by our actions instead of being heavily affected by market volatility. We can also consider the volatility scaling as normalizing rewards from different contracts to the same scale. Because our dataset consists of 50 futures contracts with different price ranges, we need to normalize different rewards to the same scale for training and for portfolio construction. μ is a fixed number per contract at each trade, and we set it to 1. Note that our transaction cost term also includes a price term, p_{t-1}. This is necessary because again we work with many contracts, each of which has a different cost, so we represent transaction cost as a fraction of traded value. The constant, basis point (bp), is the cost rate and 1 bp = 0.0001. As an example, if the cost rate is 1 bp, we need to pay \$0.10 to buy one unit of a contract priced at \$1,000.

We define $r_t = p_t - p_{t-1}$, and this expression represents additive profits. Additive profits are often used if a fixed number of shares or contracts is traded at each time. If we want to trade a fraction of our accumulated wealth at each time, multiplicative profits should be used, and $r_t = p_t/p_{t-1} - 1$. In this case, we also need to change the expression of R_t because R_t represents the percentage of our wealth. The exact form is given by Moody et al. (1998). We stick to additive profits in our work because logarithmic transformation needs to be done for multiplicative profits to have the cumulative rewards required by the RL setup, but logarithmic transformation penalizes large wealth growth.

RL Algorithms

DQN. By adopting a neural network, a DQN approximates the state-action value function (Q function) to estimate how good it is for the agent to perform a given action in a given state. Suppose our Q function is parameterized by some θ. We minimize the mean squared error between the current and target Q to derive the optimal state-action value function:

$$\begin{aligned} L(\theta) &= \mathbb{E}[(Q_\theta(S,A) - Q'_\theta(S,A))^2] \\ Q'_\theta(S_t, A_t) &= r + \gamma\ \text{argmax}_{A_{t+1}} Q_\theta(S_{t+1}, A_{t+1}) \end{aligned} \quad (5)$$

where $L(\theta)$ is the objective function. A problem is that the training of a vanilla DQN is not stable and suffers from variability. Many improvements have been made to stabilize the training process, and we adopt the following three strategies to improve the training in our work: fixed Q-targets (Van Hasselt, Guez, and Silver 2016), double DQN (Hasselt 2010), and dueling DQN (Wang et al. 2016). Fixed Q-targets and double DQN are used to reduce policy variances and to solve the problem of *chasing tails* by using a separate network to produce target values. Dueling DQNs separate the Q-value into state value and the advantage of each action. The benefit of doing this is that the value stream has more updates and we receive a better representation of the state values.

PG. The PG aims to maximize the expected cumulative rewards by optimizing the policy directly. If we use a neural network with parameters θ to represent the policy, $\pi_\theta(A|S)$, we can generate a trajectory $\tau = [S_0, A_0, R_1, S_1, \ldots, S_t, A_t, \ldots]$ from the environment and obtain a sequence of rewards. We maximize the expected cumulative rewards $J(\theta)$ by using gradient ascent to adjust θ:

$$\begin{aligned} J(\theta) &= \mathbb{E}\left[\sum_{t=0}^{T-1} R_{t+1}|\pi_\theta\right] \\ \nabla_\theta J(\theta) &= \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(A_t|S_t) G_t \end{aligned} \quad (6)$$

where G_t is defined in Equation 2. Compared with DQN, PG learns a policy directly and can output a probability distribution over actions. This is useful if we want to design stochastic policies or work with continuous action spaces. However, the training of PG uses Monte Carlo methods to sample trajectories from the environment, and the update is done only when the episode finishes. This often results in slow training, and it can get stuck at a (suboptimal) local maximum.

A2C. The A2C is proposed to solve the training problem of PG by updating the policy in real time. It consists of two models: One is an actor network that outputs the policy, and the other is a critic network that measures how good the chosen action is in the given state. We can update the policy network $\pi(A|S,\theta)$ by maximizing the objective function:

$$J(\theta) = \mathbb{E}[\log \pi(A|S,\theta) A_{adv}(S,A)] \tag{7}$$

where $A_{adv}(S, A)$ is the advantage function defined as

$$A_{adv}(S_t, A_t) = R_t + \gamma V(S_{t+1}|w) - V(S_t|w) \tag{8}$$

To calculate advantages, we use another network, the critic network, with parameters w to model the state value function $V(s|w)$, and we can update the critic network using gradient descent to minimize the temporal difference error:

$$J(w) = (R_t + \gamma V(S_{t+1}|w) - V(S_t|w))^2 \tag{9}$$

The A2C is most useful if we are interested in continuous action spaces because we reduce the policy variance by using the advantage function and update the policy in real time. The training of A2C can be done synchronously or asynchronously (A3C). In this article, we adopt the synchronous approach and execute agents in parallel on multiple environments.

EXPERIMENTS

Description of Dataset

We use data on 50 ratio-adjusted continuous futures contracts from the CLC Database (2019). Our dataset ranges from 2005 to 2019 and consists of a variety of asset classes, including commodity, equity index, fixed income, and FX. A full breakdown of the dataset can be found in Appendix A. We retrain our model at every five years, using all data available up to that point to optimize the parameters. Model parameters are then fixed for the next five years to produce out-of-sample results. In total, our testing period is from 2011 to 2019.

Baseline Algorithms

We compare our methods to the following baseline models including classical time-series momentum strategies:

- Long only
- Sign(R) (Moskowitz, Ooi, and Pedersen 2012; Lim, Zohren, and Roberts 2019):

$$A_t = \text{sign}(r_{t-252:t}) \tag{10}$$

- MACD signal (Baz et al. 2015)

$$A_t = \phi(\text{MACD}_t)$$
$$\phi(\text{MACD}) = \frac{\text{MACD} \exp(-\text{MACD}^2/4)}{0.89} \tag{11}$$

where MACD_t is defined in Equation 3. We can also take multiple signals with different time scales and average them to give a final indicator:

$$\widetilde{\text{MACD}}_t = \sum_k \text{MACD}_t(S_k, L_k) \tag{12}$$

where S_k and L_k define short and long time scales, namely $S_k \in \{8, 16, 32\}$ and $L_k \in \{24, 48, 96\}$, as done by Lim, Zohren, and Roberts (2019).

We compare the aforementioned baseline models with our RL algorithms, DQN, PG, and A2C. DQN and PG have discrete action spaces $\{-1, 0, 1\}$, and A2C has a continuous action space $[-1, 1]$.

Training Schemes for RL

In our work, we use long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) neural networks to model both actor and critic networks. LSTMs are traditionally used in natural language processing, but many recent works have applied them to financial time series (Di Persio and Honchar 2016; Bao, Yue,

EXHIBIT 1
Values of Hyperparameters for Different RL Algorithms

Model	α_{critic}	α_{actor}	Optimizer	Batch Size	γ	bp	Memory Size	τ
DQN	0.0001	–	Adam	64	0.3	0.0020	5,000	1,000
PG	–	0.0001	Adam	–	0.3	0.0020	–	–
A2C	0.001	0.0001	Adam	128	0.3	0.0020	–	–

and Rao 2017; Tsantekidis et al. 2017b; Fischer and Krauss 2018). In particular, the work of Lim, Zohren, and Roberts (2019) showed that LSTMs deliver superior performance on modeling daily financial data. We use two-layer LSTM networks with 64 and 32 units in all models, and leaky rectifying linear units (Leaky-ReLU) (Maas, Hannun, and Ng 2013) are used as activation functions. Because our dataset consists of different asset classes, we train a separate model for each asset class. It is a common practice to train models by grouping contracts within the same asset class, and we find it also improves our performance.

We list the value of hyperparameters for different RL algorithms in Exhibit 1. We denote the learning rates for critic and actor networks as α_{critic} and α_{actor}. The Adam optimizer (Kingma and Ba 2015) is used for training all networks, and batch size means the size of the minibatch used is in a gradient descent. As previously introduced, γ is the discounting factor and bp is the cost rate used in training. We can treat bp as a regularizing term; a large bp penalizes turnovers and lets agents maintain current trade positions. The memory size shows the size of the buffer for experience replay, and we update the parameters of our target network in DQN every τ steps.

Procedures for Controlling Overfitting

Because overfitting to backtest data can be problematic in many application domains, especially in finance, we adopt the following procedures to overcome it. The most common cause of overfitting is a poor ratio of training samples to model parameters. Although there is no universal guarantee that increasing the number of samples will resolve the overfitting issue, we find that expanding the dataset to allow training on a cross section of futures contracts, not just on a single contract, greatly helps to reduce the problem in our case. Furthermore, we intentionally design our networks to have a smaller number of free parameters whenever possible. Our models consist of two hidden layers, and each layer has a relatively small number of neurons. The resulting network has fewer parameters, which makes it less likely to overfit.

In addition, we use dropout (Hinton et al. 2016), a regularization method commonly employed in deep learning. Because our networks are small, we did not notice significant improvements, but we still recommend the usage of dropout following the general advice of selecting the least flexible model that achieves comparable cross-validation performance. This is also in line with other works (Zhang, Zohren, and Roberts 2018; Lim, Zohren, and Roberts 2019) that show improved results using dropout in financial applications, especially when handling complex network architectures.

Finally, we use 10% of any given training data as a separate cross-validation set to optimize hyperparameters. We monitor the validation performance and use early stopping with 20 epochs to help optimal model selection. Any hyperparameter optimization is done on the cross-validation set, leaving the test set for the final test performance evaluation. This reduces the leakage of the test data, another cause of overfitting.

Experimental Results

We test both baseline models and our methods between 2011 and 2019, and we calculate the trade returns net of transaction costs as in Equation 4 for each contract. We then form a simple portfolio by giving equal weights to each contract, and the trade return of a portfolio is

$$R_t^{\text{port}} = \frac{1}{N}\sum_{i=1}^{N} R_t^i \tag{13}$$

where N represents the number of contracts considered, and R_t^i is the trade return for contract i at time t. We evaluate the performance of this portfolio using the

following metrics, as suggested by Lim, Zohren, and Roberts (2019):

1. $E(R)$: annualized expected trade returns
2. std(R): annualized standard deviation of trade returns
3. Downside deviation (DD): annualized standard deviation of trade returns that are negative, also known as downside risk
4. Sharpe: annualized Sharpe ratio ($E(R)$/std(R))
5. Sortino: a variant of the Sharpe ratio that uses downside deviation as risk measures ($E(R)$/DD)
6. MDD: maximum drawdown shows the maximum observed loss from any peak of a portfolio to the trough
7. Calmar: the Calmar ratio compares the expected annual rate of return with MDD; in general, the higher the ratio, the better the performance of the portfolio
8. % +ve Returns: percentage of positive trade returns
9. $\frac{\text{Ave. P}}{\text{Ave. L}}$: the ratio between positive and negative trade returns.

We present our results in Exhibit 2, in which an additional layer of portfolio-level volatility scaling is applied for each model. This brings the volatility of different methods to the same target, so we can directly compare metrics such as expected and cumulative trade returns. We also include the results without this volatility scaling for reference in Exhibit B1 in Appendix B. Exhibit 2 is split into five parts based on different asset classes. The results show the performance of a portfolio by only using contracts from that specific asset class. An exception is the last part of the exhibit, in which we form a portfolio using all contracts from our dataset. The cumulative trade returns for different models and asset classes are presented in Exhibit 3.

We can see that RL algorithms deliver better performance for most asset classes except for the equity index, for which a long-only strategy is better. This can be explained by the fact that most equity indexes were dominated by large upward trends in the testing period. Similarly, fixed incomes had a growing trend until 2016 and then entered currently ongoing consolidation periods. Arguably, the most reasonable strategy in

EXHIBIT 2

Experiment Results for Portfolio-Level Volatility Targeting

	E(R)	Std(R)	DD	Sharpe	Sortino	MDD	Calmar	% of + Ret	Ave. P / Ave. L
Commodity									
Long	–0.710	0.979	0.604	–0.726	–1.177	0.350	–0.140	0.473	0.989
Sign(R)	0.347	0.980	0.572	0.354	0.606	0.116	0.119	0.494	1.084
MACD	–0.171	0.978	0.584	–0.175	–0.293	0.190	–0.060	0.486	1.026
DQN	**0.703**	0.973	**0.552**	**0.723**	**1.275**	**0.066**	**0.501**	**0.498**	**1.135**
PG	0.062	0.982	0.585	0.063	0.106	0.039	0.023	0.495	1.029
A2C	0.223	0.955	0.559	0.234	0.399	0.141	0.091	0.487	1.093
Equity Index									
Long	**0.668**	0.970	0.606	**0.688**	**1.102**	0.132	**0.509**	**0.542**	0.948
Sign(R)	0.228	0.966	0.610	0.236	0.374	0.344	0.077	0.528	0.930
MACD	0.016	0.962	0.618	0.017	0.027	0.311	0.006	0.519	0.927
DQN	0.629	0.970	0.606	0.648	1.038	0.161	0.381	0.541	0.944
PG	0.432	0.967	0.605	0.447	0.714	0.242	0.185	0.529	0.960
A2C	0.473	0.929	**0.593**	0.510	0.798	**0.124**	0.328	0.533	**0.962**
Fixed Income									
Long	0.680	0.975	0.576	0.698	1.180	**0.061**	0.444	0.515	1.054
Sign(R)	0.214	0.972	0.592	0.221	0.363	0.080	0.083	0.504	1.019
MACD	0.219	0.967	0.579	0.228	0.380	0.065	0.123	0.486	**1.101**
DQN	**0.908**	0.972	**0.562**	**0.935**	**1.617**	0.062	**0.543**	0.515	1.098
PG	0.705	0.974	0.576	0.724	1.225	**0.061**	0.436	**0.517**	1.052
A2C	0.699	0.979	0.582	0.714	1.203	0.067	0.408	**0.517**	1.048

(continued)

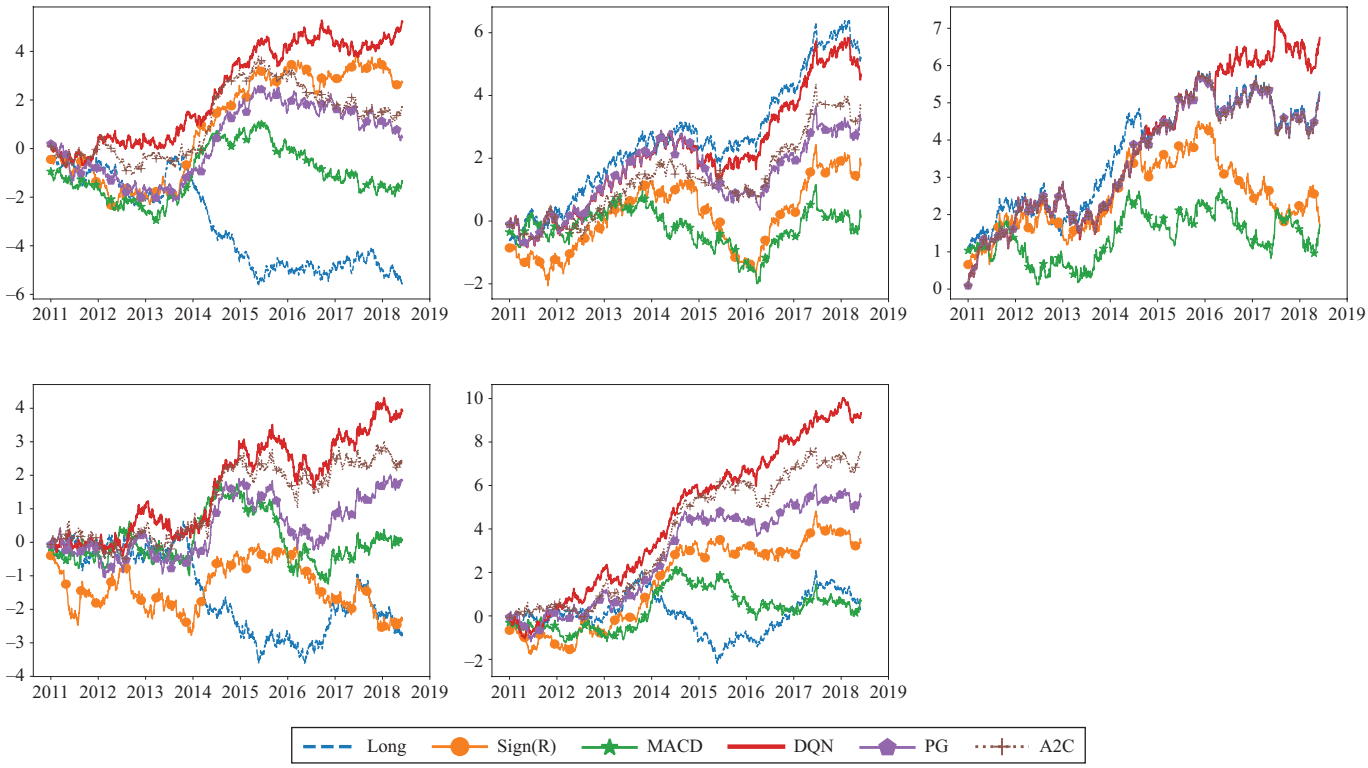
EXHIBIT 2 *(continued)*

Experiment Results for Portfolio-Level Volatility Targeting

	E(R)	Std(R)	DD	Sharpe	Sortino	MDD	Calmar	% of + Ret	Ave. P / Ave. L
FX									
Long	–0.344	0.973	0.583	–0.353	–0.590	0.423	–0.097	0.491	0.979
Sign(R)	–0.297	0.973	0.592	–0.306	–0.502	0.434	–0.111	0.499	0.954
MACD	0.006	0.970	0.582	0.007	0.011	0.329	0.002	0.493	1.029
DQN	**0.528**	0.967	**0.553**	**0.546**	**0.955**	0.183	**0.313**	**0.510**	**1.051**
PG	0.248	0.967	0.566	0.257	0.438	0.240	0.124	0.506	1.021
A2C	0.316	0.963	0.563	0.328	0.561	**0.165**	0.201	0.507	1.026
All									
Long	0.055	0.975	0.598	0.058	0.092	0.071	0.013	0.520	0.933
Sign(R)	0.429	0.972	0.582	0.441	0.737	0.038	0.201	0.510	1.031
MACD	0.089	0.978	0.582	0.091	0.153	0.008	0.035	0.493	1.043
DQN	**1.258**	0.976	**0.567**	**1.288**	**2.220**	**0.002**	**1.025**	**0.543**	**1.043**
PG	0.740	0.980	0.593	0.754	1.247	0.012	0.480	0.533	0.990
A2C	1.024	0.975	0.573	1.050	1.785	0.007	0.685	0.538	1.021

EXHIBIT 3

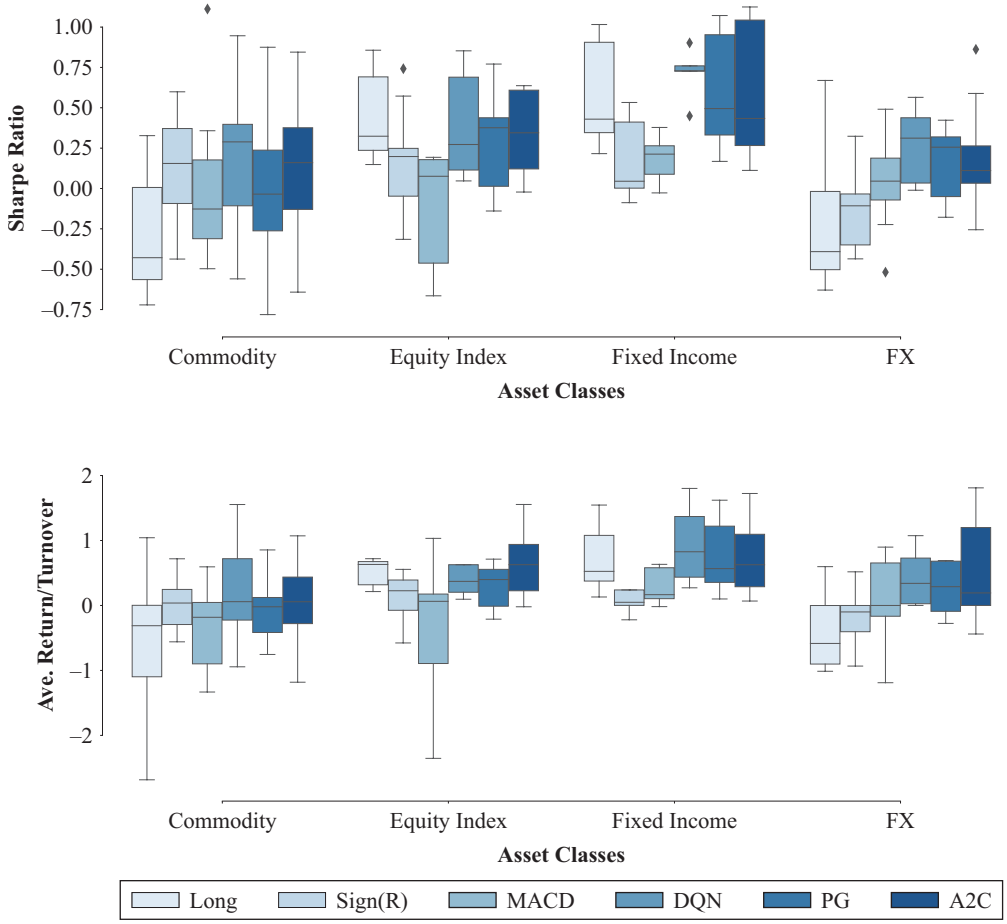
Cumulative Trade Returns



Note: First row: commodity, equity index, and fixed income; second row: FX and the portfolio of using all contracts.

Exhibit 4

Sharpe Ratio (top) and Average Trade Return per Turnover (bottom) for Individual Contracts



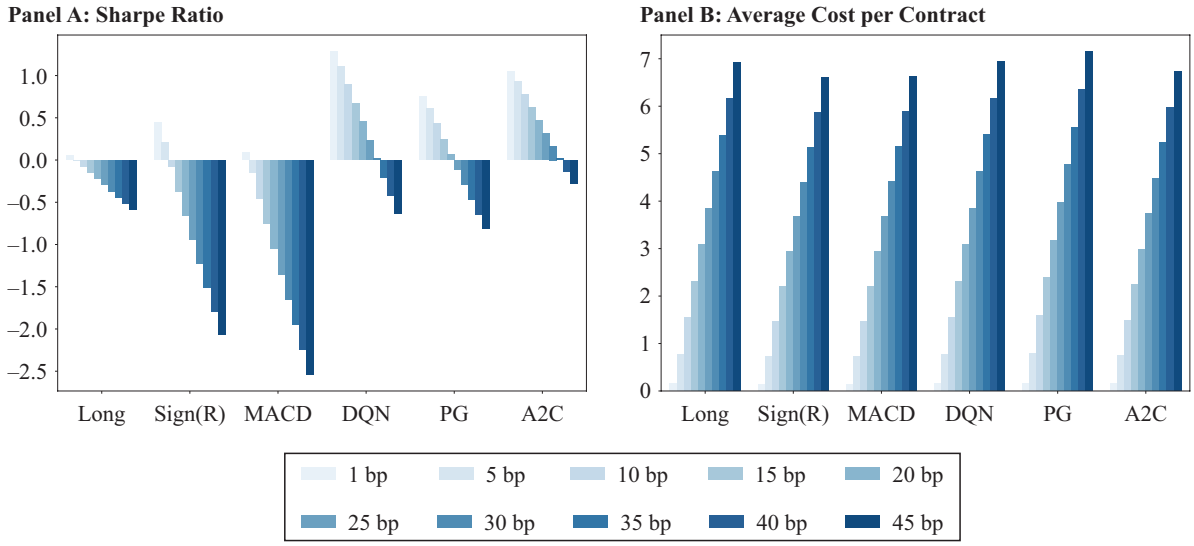
these cases might be to hold our positions if large trends persist. However, the long-only strategy performs the worst in commodity and FX markets, in which prices are more volatile. RL algorithms perform better in these markets by being able to go long or short at reasonable time points. Overall, DQN obtains the best performance among all models, and the second best is the A2C approach. We investigate the cause of this observation and find that A2C generates larger turnovers, leading to smaller average returns per turnover, as shown in Exhibit 4.

We also investigate the performance of our methods under different transaction costs. In Panel A of Exhibit 5, we plot the annualized Sharpe ratio for the portfolio using all contracts at different cost rates. We can see that RL algorithms can tolerate larger cost rates; in particular, DQN and A2C can still generate positive profits with the cost rate at 25 bp. To understand how cost rates (bp) translate into monetary values, we plot the average cost per contract in Panel B of Exhibit 5, and we see that 25 bp represents roughly $3.50 per contract. This is a realistic cost for a retail trader to pay, but institutional traders have a different fee structure based on trading volumes and often have cheaper cost. In any case, this shows the validity of our methods in a realistic setup.

The performance of a portfolio is generally better than the performance of an individual contract because risks are diversified across a range of assets, so the return per risk is higher. To investigate the raw quality of our

EXHIBIT 5

Sharpe Ratio and Average Cost per Contract under Different Cost Rates



methods, we investigate the performance of individual contracts. We use the boxplots in Exhibit 4 to present the annualized Sharpe ratio and average trade return per turnover for each futures contract. Overall, these results reinforce our previous findings that RL algorithms generally work better, and the performance of our method is not driven by a single contract that shows superior performance, reassuring us about the consistency of our model.

CONCLUSION

We adopt RL algorithms to learn trading strategies for continuous futures contracts. We discuss the connection between modern portfolio theory and the RL reward hypothesis and show that they are equivalent if a linear utility function is used. Our analysis focuses on three RL algorithms, namely DQN, PG, and A2C, and investigates both discrete and continuous action spaces. We use features from time-series momentum and technical indicators to form state representations. In addition, volatility scaling is introduced to improve reward functions. We test our methods on 50 liquid futures contracts from 2011 to 2019, and our results show that RL algorithms outperform baseline models and deliver profits even under heavy transaction costs.

In continuations of this work, we would like to investigate different forms of utility functions. In practice, an investor is often risk averse, and the objective is to maximize a risk-adjusted performance function such as the Sharpe ratio, leading to a concave utility function. As suggested by Huang (2018), we can resort to distributional RL (Bellemare, Dabney, and Munos 2017) to obtain the entire distribution over $Q(s, a)$ instead of the expected Q-value. Once the distribution is learned, we can choose actions with the highest expected Q-value and the lowest standard deviation of it, maximizing the Sharpe ratio. We can also extend our methods to portfolio optimization by modifying the action spaces to give weights of individual contracts in a portfolio. We can incorporate the reward functions with mean–variance portfolio theory (Markowitz 1952) to deliver a reasonable expected trade return with minimal volatility.

Appendix A

Our dataset consists of 50 futures contracts, and there are 25 commodity contracts, 11 equity index contracts, 5 fixed income contracts, and 9 forex contracts. A detailed description of each contract is given in Exhibit A1.

Exhibit A1

Contract Descriptions

Ticker	Contract Details	Ticker	Contract Details
Commodities		**Equity Indexes**	
CC	COCOA	CA	CAC40 INDEX
DA	MILK III, Comp	ER	RUSSELL 2000, MINI
GI	GOLDMAN SAKS C. I.	ES	S&P 500, MINI
JO	ORANGE JUICE	LX	FTSE 100 INDEX
KC	COFFEE	MD	S&P 400 (Mini Electronic)
KW	WHEAT, KC	SC	S&P 500, Composite
LB	LUMBER	SP	S&P 500, Day Session
NR	ROUGH RICE	XU	DOW JONES EUROSTOXX50
SB	SUGAR #11	XX	DOW JONES STOXX 50
ZA	PALLADIUM, Electronic	YM	Mini Dow Jones ($5.00)
ZC	CORN, Electronic	**Fixed Incomes**	
ZF	FEEDER CATTLE, Electronic	DT	EURO BOND (BUND)
ZG	GOLD, Electronic	FB	T-NOTE, 5-Year Composite
ZH	HEATING OIL, Electronic	TY	T-NOTE, 10-Year Composite
ZI	SILVER, Electronic	UB	EURO BOBL
ZK	COPPER, Electronic	US	T-BONDS, Composite
ZL	SOYBEAN OIL, Electronic	**Foreign Exchange**	
ZN	NATURAL GAS, Electronic	AN	AUSTRALIAN, Day Session
ZO	OATS, Electronic	BN	BRITISH POUND, Composite
ZP	PLATINUM, Electronic	CN	CANADIAN, Composite
ZR	ROUGH RICE, Electronic	DX	US DOLLAR INDEX
ZT	LIVE CATTLE, Electronic	FN	EURO, Composite
ZU	CRUDE OIL, Electronic	JN	JAPANESE YEN, Composite
ZW	WHEAT, Electronic	MP	MEXICAN PESO
ZZ	LEAN HOGS, Electronic	NK	NIKKEI INDEX
		SN	SWISS FRANC, Composite

Appendix B

Exhibit B1 presents the performance metrics for portfolios without the additional layer of volatility scaling.

Exhibit B1

Experiment Results for the Raw Signal

	E(R)	Std(R)	DD	Sharpe	Sortino	MDD	Calmar	% of + Ret	Ave. P / Ave. L
Commodity									
Long	–0.298	0.412	0.258	–0.723	–1.152	0.248	–0.130	0.473	0.987
Sign(R)	0.101	0.312	0.185	0.325	0.548	0.082	0.115	0.494	1.081
MACD	–0.039	0.227	0.136	–0.174	–0.290	0.132	–0.059	0.486	1.024
DQN	0.187	0.301	0.173	0.623	1.085	0.039	0.413	0.498	1.119
PG	0.013	0.287	0.172	0.047	0.078	0.072	0.017	0.495	1.026
A2C	0.072	0.163	0.098	0.440	0.729	0.099	0.161	0.487	1.151

(continued)

EXHIBIT B1 *(continued)*

Experiment Results for the Raw Signal

	E(R)	Std(R)	DD	Sharpe	Sortino	MDD	Calmar	% of + Ret	Ave. P / Ave. L
Equity Indexes									
Long	0.504	0.928	0.606	0.543	0.831	0.127	0.466	0.541	0.928
Sign(R)	0.168	0.799	0.526	0.211	0.319	0.299	0.075	0.528	0.928
MACD	–0.068	0.586	0.385	–0.117	–0.178	0.351	–0.041	0.519	0.904
DQN	0.461	0.933	0.611	0.494	0.754	0.170	0.308	0.541	0.922
PG	0.320	0.875	0.574	0.366	0.558	0.211	0.183	0.529	0.949
A2C	0.293	0.629	0.427	0.466	0.686	0.193	0.214	0.533	0.965
Fixed Income									
Long	0.605	0.939	0.561	0.645	1.081	0.108	0.455	0.515	1.048
Sign(R)	0.189	0.795	0.496	0.237	0.381	0.165	0.103	0.504	1.024
MACD	0.136	0.609	0.367	0.224	0.371	0.124	0.131	0.485	1.102
DQN	0.734	0.862	0.508	0.851	1.445	0.118	0.469	0.515	1.086
PG	0.624	0.938	0.561	0.665	1.113	0.109	0.443	0.517	1.043
A2C	0.852	1.345	0.806	0.633	1.057	0.128	0.397	0.517	1.039
Foreign Exchange									
Long	–0.198	0.472	0.285	–0.420	–0.696	0.219	–0.101	0.491	0.966
Sign(R)	–0.113	0.551	0.341	–0.207	–0.332	0.170	–0.071	0.499	0.968
MACD	0.016	0.424	0.259	0.037	0.061	0.156	0.016	0.493	1.034
DQN	0.272	0.487	0.280	0.560	0.972	0.085	0.326	0.510	1.058
PG	0.157	0.533	0.312	0.295	0.503	0.098	0.148	0.506	1.029
A2C	0.159	0.455	0.267	0.349	0.592	0.081	0.193	0.507	1.034
All									
Long	–0.013	0.363	0.230	–0.036	–0.057	0.037	–0.009	0.519	0.919
Sign(R)	0.086	0.296	0.186	0.291	0.461	0.016	0.142	0.510	1.008
MACD	–0.018	0.230	0.143	–0.080	–0.129	0.026	–0.029	0.493	1.013
DQN	0.318	0.252	0.150	1.258	2.111	0.008	1.023	0.543	1.041
PG	0.168	0.279	0.174	0.602	0.968	0.011	0.373	0.533	0.968
A2C	0.214	0.221	0.134	0.969	1.601	0.009	0.672	0.538	1.014

ACKNOWLEDGMENT

The authors would like to thank Bryan Lim, Vu Nguyen, Anthony Ledford, and members of Machine Learning Research Group at the University of Oxford for their helpful comments. We are most grateful to the Oxford-Man Institute of Quantitative Finance, which provided the Pinnacle dataset and computing facilities.

REFERENCES

Arrow, K. J. "The Theory of Risk Aversion." In *Essays in the Theory of Risk-Bearing*, pp. 90–120. Chicago: Markham, 1971.

Baltas, N., and R. Kosowski. "Demystifying Time-Series Momentum Strategies: Volatility Estimators, Trading Rules and Pairwise Correlations." *Trading Rules and Pairwise Correlations*, May 8, 2017.

Bao, W., J. Yue, and Y. Rao. 2017. "A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long–Short Term Memory." *PLOS One* 12 (7): e0180944.

Baz, J., N. Granger, C. R. Harvey, N. Le Roux, and S. Rattray. "Dissecting Investment Strategies in the Cross Section and Time Series." SSRN 2695101, 2015.

Bekiros, S. D. 2010. "Heterogeneous Trading Strategies with Adaptive Fuzzy Actor–Critic Reinforcement Learning: A Behavioral Approach." *Journal of Economic Dynamics and Control* 34 (6): 1153–1170.

Bellemare, M. G., W. Dabney, and R. Munos. 2017. "A Distributional Perspective on Reinforcement Learning." *Proceedings of the 34th International Conference on Machine Learning* 70: 449–458.

Bertoluzzo, F., and M. Corazza. 2012. "Testing Different Reinforcement Learning Configurations for Financial Trading: Introduction and Applications." *Procedia Economics and Finance* 3: 68–77.

Chan, E. *Quantitative Trading: How to Build Your Own Algorithmic Trading Business*, vol. 430. Hoboken: John Wiley & Sons, 2009.

——. *Algorithmic Trading: Winning Strategies and Their Rationale*, vol. 625. Hoboken: John Wiley & Sons, 2013.

CLC Database. Pinnacle Data Corp, 2019, https://pinnacledata2.com/clc.html.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. "Natural Language Processing (Almost) from Scratch." *Journal of Machine Learning Research* 12 (Aug): 2493–2537.

Deng, Y., F. Bao, Y. Kong, Z. Ren, and Q. Dai. 2016. "Deep Direct Reinforcement Learning for Financial Signal Representation and Trading." *IEEE Transactions on Neural Networks and Learning Systems* 28 (3): 653–664.

Di Persio, L., and O. Honchar. 2016. "Artificial Neural Networks Architectures for Stock Price Prediction: Comparisons and Applications." *International Journal of Circuits, Systems and Signal Processing* 10: 403–413.

Fischer, T., and C. Krauss. 2018. "Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions." *European Journal of Operational Research* 270 (2): 654–669.

Fischer, T. G. "Reinforcement Learning in Financial Markets—A Survey." Technical report, FAU Discussion Papers in Economics, 2018.

Graham, B., and D. L. F. Dodd. *Security Analysis*. New York: McGraw-Hill, 1934.

Harvey, C. R., E. Hoyle, R. Korgaonkar, S. Rattray, M. Sargaison, and O. Van Hemert. 2018. "The Impact of Volatility Targeting." *The Journal of Portfolio Management* 45 (1): 14–33.

Hasselt, H. V. "Double Q-Learning." In *Advances in Neural Information Processing Systems*, pp. 2613–2621. Cambridge, MA: MIT Press, 2010.

Hinton, G. E., A. Krizhevsky, I. Sutskever, and N. Srivastva. "System and Method for Addressing Overfitting in a Neural Network." US Patent 9,406,017, August 2, 2016.

Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–1780.

Huang, C. Y. 2018. "Financial Trading as a Game: A Deep Reinforcement Learning Approach." *arXiv* 1807.02787.

Ingersoll, J. E. *Theory of Financial Decision Making,* vol. 3. Lanham, MD; Rowman & Littlefield, 1987.

Jin, O., and H. El-Saawy. "Portfolio Management Using Reinforcement Learning." Technical report working paper, Stanford University, 2016.

Kingma, D. P., and J. Ba. "Adam: A Method for Stochastic Optimization." *Proceedings of the International Conference on Learning Representations*, 2015.

Kirk, D. E. *Optimal Control Theory: An Introduction*. North Chelmsford: MA; Courier Corporation, 2012.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. "Imagenet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, pp. 1097–1105. Cambridge, MA: MIT Press, 2012.

Li, H., C. H. Dagli, and D. Enke. 2007. "Short-Term Stock Market Timing Prediction under Reinforcement Learning Schemes." *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pp. 233–240.

Lim, B., S. Zohren, and S. Roberts. 2019. "Enhancing Time-Series Momentum Strategies Using Deep Neural Networks." *The Journal of Financial Data Science* 1 (4): 19–38.

Maas, A. L., A. Y. Hannun, and A. Y. Ng. 2013. "Rectifier Nonlinearities Improve Neural Network Acoustic Models." *ICML Workshop on Deep Learning for Audio, Speech and Language Processing* 30: 3.

Markowitz, H. 1952. "Portfolio Selection." *The Journal of Finance* 7 (1): 77–91.

Merton, R. C. 1969. "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case." *The Review of Economics and Statistics* 51 (3): 247–257.

Metropolis, N., and S. Ulam. 1949. "The Monte Carlo Method." *Journal of the American Statistical Association* 44 (247): 335–341.

Michaud, R. O. 1989. "The Markowitz Optimization Enigma: Is 'Optimized' Optimal?" *Financial Analysts Journal* 45 (1): 31–42.

Mnih, V., A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. 2016. "Asynchronous Methods for Deep Reinforcement Learning." In *International Conference on Machine Learning* 48: 1928–1937.

Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. "Playing Atari with Deep Reinforcement Learning." NIPS Deep Learning Workshop, 2013.

Moody, J., and M. Saffell. 2001. "Learning to Trade via Direct Reinforcement. *IEEE Transactions on Neural Networks* 12 (4): 875–889.

Moody, J., L. Wu, Y. Liao, and M. Saffell. 1998. "Performance Functions and Reinforcement Learning for Trading Systems and Portfolios." *Journal of Forecasting* 17 (5–6): 441–470.

Moskowitz, T. J., Y. H. Ooi, and L. H. Pedersen. 2012. "Time Series Momentum." *Journal of Financial Economics* 104 (2): 228–250.

Murphy, J. J. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methos and Applications*. Paramus, NJ: New York Institute of Finance, 1999.

O'Neil, W. J. *How to Make Money in Stocks: A Winning System in Good Times and Bad*, 4th ed. New York: McGraw-Hill, 2009.

Pratt, J. W. "Risk Aversion in the Small and in the Large." In *Uncertainty in Economics*, pp. 59–79. Elsevier, 1978.

Ritter, G. 2017. "Machine Learning for Trading." August 8, 2017. Available at SSRN: https://ssrn.com/abstract=3015609 or http://dx.doi.org/10.2139/ssrn.3015609.

Rohrbach, J., S. Suremann, and J. Osterrieder. 2017. "Momentum and Trend Following Trading Strategies for Currencies Revisited—Combining Academia and Industry." SSRN 2949379.

Schwager, J. D. *A Complete Guide to the Futures Market: Technical Analysis, Trading Systems, Fundamental Analysis, Options, Spreads, and Trading Principles*. Hoboken: John Wiley & Sons, 2017.

Sirignano, J., and R. Cont. 2019. "Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning." *Quantitative Finance* 19 (9): 1449–1459.

Sutton, R. S., and A. G. Barto. *Introduction to Reinforcement Learning*, vol. 2. Cambridge, MA: MIT Press, 1998.

Tan, Z., C. Quek, and P. Y. Cheng. 2011. "Stock Trading with Cycles: A Financial Application of Anfis and Reinforcement Learning." *Expert Systems with Applications* 38 (5): 4741–4755.

Tsantekidis, A., N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis. 2017a. "Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks." *2017 IEEE 19th Conference on Business Informatics (CBI)* 1: 7–12.

——. "Using Deep Learning to Detect Price Change Indications in Financial Markets." 2017b. *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2511–2515.

Van Hasselt, H., A. Guez, and D. Silver. "Deep Reinforcement Learning with Double Q-Learning." *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Wang, Z., T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas. 2016. "Dueling Network Architectures for Deep Reinforcement Learning." In *Proceedings of the 33rd International Conference on International Conference on Machine Learning* 48: 1995–2003.

Wilder, J. W. "New Concepts in Technical Trading Systems." Trend Research, 1978.

Williams, R. J. 1992. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning." *Machine Learning* 8 (3–4): 229–256.

Xiong, Z., X. Y. Liu, S. Zhong, and A. Walid. 2018. "Practical Deep Reinforcement Learning Approach for Stock Trading." *arXiv* 1811.07522.

Zhang, Z., S. Zohren, and S. Roberts. "Bdlob: Bayesian Deep Convolutional Neural Networks for Limit Order Books." Presented at Third Workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada, 2018.

——. 2019a. "Deeplob: Deep Convolutional Neural Networks for Limit Order Books." *IEEE Transactions on Signal Processing* 67 (11): 3001–3012.

——. "Extending Deep Learning Models for Limit Order Books to Quantile Regression." *Proceedings of Time Series Workshop of the 36th International Conference on Machine Learning*, 2019b.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

The Impact of Volatility Targeting

Campbell R. Harvey, Edward Hoyle, Russell Korgaonkar, Sandy Rattray, Matthew Sargaison, and Otto Van Hemert

The Journal of Portfolio Management
https://jpm.pm-research.com/content/45/1/14

ABSTRACT: *Recent studies show that volatility-managed equity portfolios realize higher Sharpe ratios than portfolios with a constant notional exposure. The authors show that this result only holds for risk assets, such as equity and credit, and they link this finding to the so-called leverage effect for those assets. In contrast, for bonds, currencies, and commodities, the impact of volatility targeting on the Sharpe ratio is negligible. However, the impact of volatility targeting goes beyond the Sharpe ratio: It reduces the likelihood of extreme returns across all asset classes. Particularly relevant for investors, left-tail events tend to be less severe because they typically occur at times of elevated volatility, when a target-volatility portfolio has a relatively small notional exposure. We also consider the popular 60–40 equity–bond balanced portfolio and an equity–bond–credit–commodity risk parity portfolio. Volatility scaling at both the asset and portfolio level improves Sharpe ratios and reduces the likelihood of tail events.*

A Century of Evidence on Trend-Following Investing

Brian Hurst, Yao Hua Ooi, and Lasse Heje Pedersen

The Journal of Portfolio Management
https://jpm.pm-research.com/content/44/1/15

ABSTRACT: *In this article, the authors study the performance of trend-following investing across global markets since 1880, extending the existing evidence by more than 100 years using a novel data set. They find that in each decade since 1880, time-series momentum has delivered positive average returns with low correlations to traditional asset classes. Further, time-series momentum has performed well in 8 out of 10 of the largest crisis periods over the century, defined as the largest drawdowns for a 60/40 stock/bond portfolio. Lastly, the authors find that time-series momentum has performed well across different macro environments, including recessions and booms, war and peace, high- and low-interest-rate regimes, and high- and low-inflation periods.*

Enhancing Time-Series Momentum Strategies Using Deep Neural Networks

Bryan Lim, Stefan Zohren, and Stephen Roberts

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/4/19

ABSTRACT: *Although time-series momentum is a well-studied phenomenon in finance, common strategies require the explicit definition of both a trend estimator and a position sizing rule. In this article, the authors introduce deep momentum networks—a hybrid approach that injects deep learning–based trading rules into the volatility scaling framework of time-series momentum. The model also simultaneously learns both trend estimation and position sizing in a data-driven manner, with networks directly trained by optimizing the Sharpe ratio of the signal. Backtesting on a portfolio of 88 continuous futures contracts, the authors demonstrate that the Sharpe-optimized long short-term memory improved traditional methods by more than two times in the absence of transactions costs and continued outperforming when considering transaction costs up to 2–3 bps. To account for more illiquid assets, the authors also propose a turnover regularization term that trains the network to factor in costs at run-time.*

Modular Machine Learning for Model Validation: *An Application to the Fundamental Law of Active Management*

JOSEPH SIMONIAN

JOSEPH SIMONIAN
is the founder and CIO of Autonomous Investment Technologies LLC, in Newton, MA.
jsslog15@gmail.com

KEY FINDINGS

- Implementing model validation through a set of interdependent modules that uses both traditional econometrics and data science techniques can produce robust assessments of the predictive effectiveness of investment signals in an economically intuitive manner.
- The proposed methodology, modular machine learning, also answers several practical questions that arise when applying block time series cross-validation, such as the length of the folds and the block size to use between folds.
- It is possible to reinterpret the fundamental law of active management into a model validation framework by expressing its fundamental concepts—information coefficient and breadth—using the formal language of data science.

ABSTRACT: *The author introduces a modular machine learning framework for model validation in which the output from one procedure serves as the input to another procedure within a single validation framework. A defining feature of the described methodology is the use of both traditional econometrics and data science. The author uses an econometric model in the first module to classify data in an economically intuitive way. Proceeding modules apply data science techniques to evaluate the predictive characteristics of the model components. The author applies his framework to the fundamental law of active management, a well-known formal characterization of portfolio managers' alpha generation process. In contrast to standard applications of the law, in which it has been used to evaluate a manager's existing active management process, the author recasts the law within his framework as a means to test investment signals for potential use, individually or collectively, in a manager's investment process. To illustrate how this application works, the author provides an example using the well-known Fama–French factors as test signals.*

TOPICS: ***Statistical methods, simulations, big data/machine learning****

The models that drive the alpha engines of quantitative investment firms generally consist of a multitude of investment signals. Ideally, each signal within a given alpha engine will be a sufficiently unique driver of excess returns, in the sense that each signal will provide information that is both

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

predictively useful and complementary to the other signals feeding into an investment process. As such, model validation is a critical component of quantitative investing. The construction of any model, whether theoretical or applied, requires a balancing of informative power and simplicity. In the context of investment models, these considerations respectively manifest themselves in predictive efficacy and signal variety.

The latter considerations are also central to the well-known theoretical framework known as The Fundamental Law of Active Management (FL) (Grinold 1989; Grinold and Kahn 2000), which defines an investment strategy's information ratio (IR) using the formula $IR = IC \times \sqrt{BR}$. The term IC stands for the expected *information coefficient*, which (broadly construed) represents the predictive efficacy of the forecasts or signals used in an active investment strategy. The IC is a straightforward measure of the informative power of an investment model. Formally, the IC is taken to be the *ex ante* correlation that a signal or forecast has with realized returns. The term BR stands for *breadth*, which has alternatively been interpreted as the number of independent forecasts, signals, or positions in a given investment strategy. Although the interpretation of BR within the context of model construction is not as straightforward as the interpretation of the IC, BR is, in fact, related to the notion of model simplicity. If a given signal is not an independent source of information relative to any other signal in a model, then either it or the signal to which it is being compared is redundant and should be discarded.

Although "active management is forecasting" (Kahn 1997) and the FL provides the basis upon which to build an elegant model validation framework, most research conducted on the FL over the years has been primarily concerned with applying it toward performance analysis.[1] In this regard, there has been a particular emphasis on developing the formal machinery of the FL so that the dispersion between its *ex ante* specified IR value and its *ex post* measurement is minimized. Although prior research on the FL has been useful in providing a better understanding of how to bring the specification of *ex ante* IRs more in line with market and portfolio management realities, it has not given us the formal means to employ the FL as way of assessing the *predictive* validity of a given investment framework. Specifically, it has not materially increased our ability to use the FL as a practical tool for assessing the robustness of the individual investment signals that form the basis of any investment strategy.

In this article, we introduce an approach to model validation that we call *modular machine learning* (MML) and use it to build a methodology that can be applied to the evaluation of investment signals within the conceptual scheme provided by the FL. Our framework is modular in two respects:

1. It is composed of independent computational components, each using the output of another as its input.[2]
2. It is characterized by the distinct role played by traditional econometric and data science methodologies.[3]

WHICH FUNDAMENTAL LAW?

Although the original FL identified the IC and BR as the essential inputs into the generation of IRs, subsequent research has argued that additional metrics are critical to the evaluation of actively managed portfolios. One metric that has been emphasized in several frameworks is the volatility of the IC, which is considered important because it is a way to incorporate the temporal variability of alpha into the calculation of the IR. One framework that incorporates IC volatility has been

[1] See, for instance, Grinold (1994); Clarke, de Silva, and Thorley (2002, 2006); Buckle (2004); Qian and Hua (2004); Zhou (2008); and Menchero (2017). Moreover, we note that some variants of the FL contain so much additional formal accoutrement that they are more rightly viewed as standalone models rather than extensions of the FL. See Ding and Martin (2017) for an example.

[2] In this respect, there is a family resemblance between MML and the idea of a *modular deep network* as described by Ghazi, Panigrahy, and Wang (2019). A modular deep network consists of independent neural networks that only communicate via one's output serving as another's input. Although our concept of modular machine learning is not tied to any particular type of algorithm, it is also defined by the dependence of one module on another for input. The trading methodology presented by Simonian et al. (2019) can be considered a precursor to the methodology presented here because it uses the random forest algorithm to produce output that feeds into a trading strategy that uses another machine learning framework, association rule learning (Agrawal, Imieliński, and Swami 1993).

[3] Another recent framework that combines econometric and machine learning methodologies is the approach called *double machine learning*. See Chernozhukov et al. (2018) for an example.

developed by Ye (2008). It is formally expressed in the following manner:

$$\text{Information ratio} = \frac{IC}{\sqrt{\frac{1}{N} + \sigma_{IC}^2}} \tag{1}$$

where σ_{IC}^2 represents the variance of the IC. Although there are various ways to express the variability of the IC, in this article we will build our framework using the formula in Equation 1 because it extends the original FL in a formally elegant manner with few underlying assumptions.

MODULAR MACHINE LEARNING

Model validation using MML is implemented through three distinct modules:

- *Subsample classification module*: An econometric model is used to classify the subsets of a time series into distinct regimes.
- *Signal quality module*: The regime-specific subsets of data from the subsample classification module are used as inputs to block time-series cross-validated regressions to derive regime-specific measures of signal quality. In the context of the FL, these would be values for the IC and IC variance.
- *Signal diversification module*: Using the measurements from the signal quality module as inputs, clustering is applied to determine the level of diversity present in a set of investment signals. In the context of the FL, the number of clusters is used as the value for BR.

MML is designed to be flexible enough to accommodate the use of different types of models and algorithms within the model validation process. Nevertheless, the modular nature of the framework ensures an inherent degree of structural cohesiveness to the model validation procedure, irrespective of the specific techniques chosen to carry it out.

A defining characteristic of MML is its blending of what Breiman (2001) calls the "data modeling culture" represented by econometrics, with the "algorithmic modeling culture" represented by data science. The data modeling culture has hitherto been the dominant culture in traditional statistical practice and is primarily concerned with providing information about the relationships that exist between input and response variables; it assumes that the data are generated by a specific stochastic process. In contrast, the algorithmic modeling culture assumes that the relationships between input and response variables are essentially too complex to uncover; consequently, it is primarily concerned with devising methods to successfully predict responses from inputs. It does not make any *a priori* assumptions on the data distributions.[4]

Our "multiculturalism" is motivated by several practical considerations. The first is the recognition that data science and traditional econometrics each have unique strengths that can prove valuable to the model validation process. Data science is directed toward prediction, testing, and pattern recognition. Thus, it offers distinct value in measuring the predictive efficacy and uniqueness of individual investment signals. In contrast, traditional econometrics has a fairly weak track record with regard to producing effective forecasting tools,[5] but it has been more successful in producing formal frameworks that provide *ex post* insight into the structure of economic data.[6]

Indeed, our use of an econometric model in the subsample classification module is motivated by the distinct advantages of using an econometrically rigorous foundation upon which to base model validation, both in terms of methodological clarity and technical utility.

First, using regime-switching models to classify time series provides the formal means for the model validation process to speak to the particular mental models of the economic and financial world that portfolio managers use when building investment strategies. A *mental model* is a general picture of the fundamental drivers of markets that helps frame investment decisions.[7] For example, a macro investor's mental model

[4] An alternative view is that statistical modeling techniques fall on a continuum rather than being represented by two distinct cultures. Simonian and Fabozzi (2019) provide an argument in support of this view in relation to financial data science.

[5] Regime-switching models have proven to have a low degree of predictive efficacy. For a data-science–driven alternative to traditional regime-switching models, see Simonian and Wu (2019).

[6] This comment should not be construed as a criticism of traditional econometrics, given that, historically, its emphasis has been on the identification and characterization of causal structures in economic and financial data.

[7] The use of mental models shares an affinity with a view advanced by Binmore (2009) that knowledge is commitment to a model. This notion, which Binmore claims is the *de facto* mode of scientific discourse, dispenses with traditional criteria for knowledge

might consider the interaction between macroeconomic factors, such as gross domestic product (GDP) and inflation, as the most relevant drivers of investment decisions. An equity investor's mental model, on the other hand, might consider the interaction of company earnings and overall market volatility. In general, most investors' mental models are not as simple as the foregoing examples. Nevertheless, the use of regime-switching models can be useful in providing a delineation of economic and financial data that is consistent with an investor's decision-making process.

The second advantage of using regime-switching models for subsample classification is that when their output is used as the input to a cross-validation procedure, the length of the folds are objectively determined by the regime-switching model employed. Although the folds belonging to distinct regimes will almost invariably be of different lengths, each subset of folds will represent specific economic and market states. For example, in a two-state regime-switching model, one portion of the folds will correspond to one regime, and another portion of the folds will correspond to the second regime. Deriving folds in this manner will allow signals to be evaluated relative to a specific regime, providing portfolio managers with insight into the effectiveness of their investment process in different market conditions. Accordingly, in the context of the FL, this regime-based cross-validation allows for the derivation of predictive IRs corresponding to different regimes, which can then be considered alone or combined through various averaging procedures.

The third advantage of using regime-switching models in model validation is in relation to the block size in block time-series cross-validation. As is well known, time series contain unique properties that distinguish them from other types of structured data. Perhaps the most fundamental is that the values in a time series at each point possess a memory of past values. Time-series cross-validation is intended to address the memory of the past that is embedded in data that exhibit chronological dependencies. It differs from standard types of cross-validation in its requirement that training data temporally precede test data. Although time-series cross-validation solves the memory problem in chronologically dependent data, overfitting can still be a problem because test samples are likely to have some memory of training samples that immediately precede them in time. As a remedy, block time-series cross-validation has been introduced.[8] This procedure is defined by the omission of some observations (blocks) that lie between training and test samples during cross-validation to reinforce the memoryless nature of the operation. When blocking is used, the question of block size naturally arises, but time-series cross-validation by itself does not provide a clear-cut way to choose a value. Traditional regime-switching models, on the other hand, can provide some guidance in this respect because they specify the order of the autoregressive process that drives regimes. This order value seems to be a natural choice for block size; it indicates the length of time beyond which the time series forgets.

We close this section by noting that our ecumenical approach to statistical analysis in MML is rooted in an instrumentalist view of economic theories and the FL in particular. In this respect, we treat the FL in the same spirit that Friedman (1953) treats economic models in general: as a means "to serve as a filing system for organizing empirical material."[9] That is, within the framework presented, IRs are simply considered the formal means to compare the predictive efficacy and distinctiveness of individual investment signals. Thus, we make no claim that we are providing a better method for uncovering "true IRs".

POPULATING THE MODULES

With the description of the methodology complete, we proceed with a concrete example that demonstrates how MML can be used to test and classify the predictive utility of investment signals. Our first step is to flesh out the schematic for MML presented in the previous section of the article by selecting specific models and algorithms to use in each module of the model validation process. Our choice of models and algorithms is primarily motivated by a desire to present a tractable example for expository purposes. In the case

such as "justification" and "truth" and instead characterizes "knowing" a proposition as acting as if the proposition were true.

[8] See Burman, Chow, and Nolan (1994), Racine (2000), Bergmeir and Benítez (2012), Roberts et al. (2017), and Bergmeir, Hyndman, and Koo (2018) for discussions on various aspects of block time-series cross-validation.

[9] See Boland (2016) for a review of Friedman's instrumentalism.

of the clustering technique known as *affinity propagation*, we have also chosen it because it does not require the prespecification of the number of clusters and so can be used as an objective measure of BR.[10]

Subsample Classification Module

Filardo's (1994) extension to Hamilton's (1989) well-known regime-switching model is used to demarcate the length and number of folds. Filardo follows Hamilton in setting up his model as an autoregressive process:

$$y_t = \mu_{S_t} + \phi_1(y_{t-1} - \mu_{S_{t-1}}) + \phi_2(y_{t-2} - \mu_{S_{t-2}}) \tag{2}$$

where y_t is a economic variable, μ_{S_t} is the intercept, s_t is a regime process, and ϕ is the autocorrelation with random shocks $\varepsilon_t \sim N(0, \sigma^2)$. Period-to-period transitions from one regime to another are expressed thus:

$$P(S_t = s_t \mid S_{t-1} = s_{t-1}) = \begin{bmatrix} p_{00,t} & p_{10,t} \\ p_{01,t} & p_{11,t} \end{bmatrix} \tag{3}$$

However, Filardo extends Hamilton's model by also allowing for exogenous regressors to influence transition probabilities:

$$p_{i,j,t} = \frac{\exp\left\{x'_{t-1}\beta_{i,j}\right\}}{1 + \exp\left\{x'_{t-1}\beta_{i,j}\right\}} \tag{4}$$

where $p_{i,j,t}$ is the probability of transitioning from regime i to regime j in period t. In our implementation, we use the year-over-year change in GDP as our endogenous regressor and the effective federal funds rate as our exogenous regressor.

Signal Quality Module

We use k-nearest neighbor regression (k-NNR) to derive the IC. Block time-series cross-validated univariate regressions are run on each regime subsample, and the predictive R^2 (score) for each is used as the IC value. Regime-specific ICs can be used in a standalone fashion or averaged to derive a global IC value. k-NNR (Fix and Hodges 1951; Cover and Hart 1967) is a nonparametric regression framework that derives predicted values for target variables in a test sample by searching the training sample for the k observations that are nearest to the predictor variable values being considered and using their average response, $r_n^{k-\mathrm{NN}}(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i,n)}(x)$, as the predicted response for the target variable. The property of nearness is defined formally using a distance measure such as Euclidean distance, $d(x_j, x_k) = \sqrt{\sum_i (x_{j,i} - x_{k,i})^2}$.

Signal Diversification Module

We derive BR using the clustering methodology known as *affinity propagation* (AP), introduced by Frey and Dueck (2007), to determine the level of diversity within a set of investment signals. To classify signals according to their statistical predictive properties, we posit a vector consisting of IC and IC variance values in various regime-specific folds. Defining BR in this manner allows us to gain a multidimensional view of signal diversity, with each measure providing information on a different aspect of predictive effectiveness. AP belongs to the family of graph-theoretic clustering techniques and is based on the concept of "message passing" (Mézard 2007) between the candidate members of a cluster that continues until each candidate is sufficiently informed to join the appropriate cluster.

AP begins by measuring the similarity $s(i, k)$ between vectors, which represents the similarity of vector k to vector i. Similarity is measured by a metric chosen by the model builder (e.g., Euclidean distance). The basic input to AP is a real-valued number $s(k, k)$, called a *preference*, for each observation. Observations with larger preference values are more likely to be selected as cluster centers, also known as *exemplars*. However, cluster selection is not only a function of preference size but also the two message-passing operations that are the essence of the AP algorithm. The first, *responsibility* $r(i, k)$, is a message that is transmitted from observation i to a candidate exemplar k and expresses the suitability of observation k as an exemplar for observation i given the suitability of other candidate exemplars. The second,

[10] In financial applications it is not always desirable to dispense with prespecification of the number of clusters. At times, it makes sense to try different clustering schemes and let economic intuition inform the number of clusters that are ultimately chosen. See Simonian and Wu (2019) for an example.

availability $a(i,k)$, works in the opposite direction and is sent from a candidate exemplar k to an observation i. Availability expresses how appropriate it would be for observation i to select observation k as its exemplar, given the existing support observation k has from other observations to serve as an exemplar.[11] The AP algorithm begins by initializing the availabilities to zero, $a(i,k)=0$, and then proceeds to compute the responsibilities using the rule

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t.\ k' \neq k} \left\{ a(i,k') + s(i,k') \right\} \tag{5}$$

The following formula then determines whether an observation is a good exemplar:

$$a(i,k) \leftarrow \min \left\{ 0, r(k,k) \ + \sum_{i' s.t.\ i' \neq \{i,k\}} \max \left\{ \left\{ 0, r(i',k) \right\} \right\} \right\} \tag{6}$$

The "self-availability" of an observation is expressed thus:

$$a(i,k) \leftarrow \sum_{i' s.t.\ i' \neq k} \max \left\{ \left\{ 0, r(i',k) \right\} \right\} \tag{7}$$

Defining BR using AP allows us to define signal independence in a more nuanced way relative to Pearson correlation because we can observe both the strength and variability of ICs across a multitude of economically meaningful folds and use these values to assess the similarity between investment signals in terms of their predictive characteristics. The overall level of statistical redundancies present in the set of signals that constitute a given investment process will be revealed upon completion of the clustering operation, when the diversity of cluster membership becomes evident.

CASE STUDY

Now that we have populated our modules, we proceed with a concrete example. We note at the outset that our definition of "investment signal" is broad and includes not only sophisticated model-driven forecasts but any piece of information that can be rendered into a time series and systematically used to inform investment decisions. In our case study, for example, we use the Fama–French factors (Fama and French 1993) and test their ability to forecast the returns of the US manufacturing sector.[12] In addition to the three Fama–French factors, we also include the one-month lagged return of the manufacturing sector as a naïve signal that functions as a basic benchmark to evaluate the predictive efficacy of our more sophisticated signals.[13] We assume that our signals are the sole constituents of a hypothetical forecasting framework that we wish to evaluate for its predictive effectiveness.

We arrive at regime-specific IRs for high and low growth regimes using Filardo's model described in the previous section and formally defined in Equations 2, 3, and 4. As mentioned previously, the regime switches are driven by the year-over-year change in real GDP, whereas our exogenous regressor is the effective federal funds rate. Thus, the model, although simple, is intended to capture the impact that monetary policy has

[11] The messages $r(i,k)$ and $a(i,k)$ can be considered log-probability ratios.

[12] Time series for the US manufacturing sector obtained from Ken French's website: (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html).

The following factor descriptions were also obtained from the foregoing website:

- $R_m - R_f$, the excess return on the market, is the value-weighted return of all CRSP firms incorporated in the United States and listed on the NYSE, AMEX, or NASDAQ that have a CRSP share code of 10 or 11 at the beginning of month t, good shares and price data at the beginning of t, and good return data for t minus the one-month Treasury bill rate.
- SMB (small minus big) is the average return on three small portfolios minus the average return on three big portfolios:

$$SMB = \frac{Small\ value + Small\ neutral + Small_growth}{3} - \frac{Big\ value + Big\ neutral + Big_growth}{3}$$

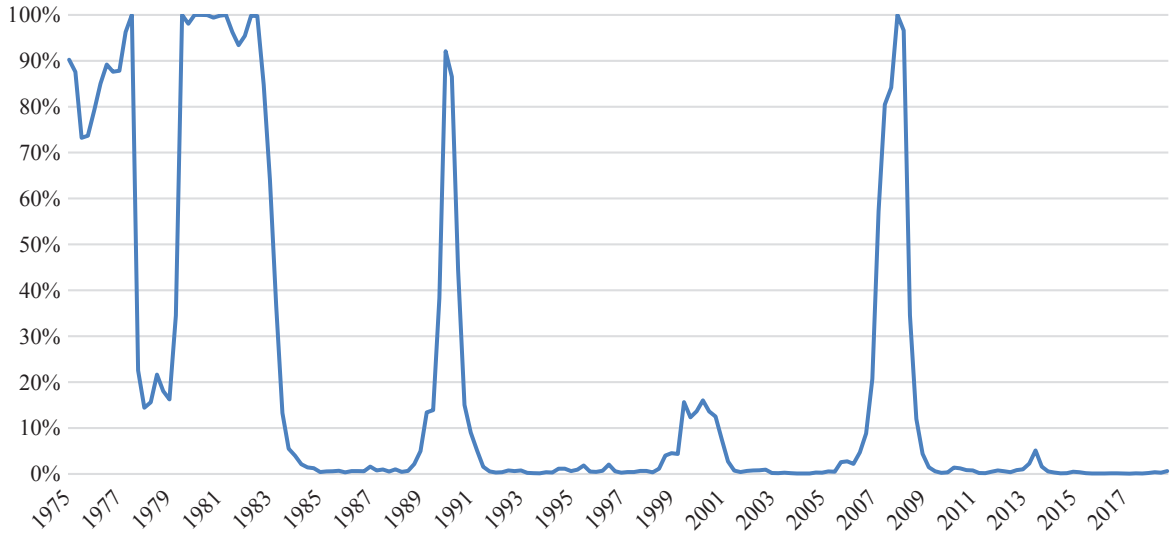
- HML (high minus low) is the average return on two value portfolios minus the average return on two growth portfolios:

$$HML = \frac{Small\ value + Big\ value}{2} - \frac{Small_growth + Big_growth}{2}$$

[13] In this respect, our model validation framework shares an affinity with the econometric concept of Granger causality (GC) (Granger 1969). GC is defined in the following manner: For a given predictor and target variable y, x *Granger-causes* y if

- x temporally precedes y and
- x provides more predictively useful information than a rival naïve predictor

EXHIBIT 1
Probability of Low Growth Regime



on driving the transition between high and low growth states. The regime classification probabilities through time are shown in Exhibit 1.

Once our states are classified, we evaluate the predictive strength of each signal over five folds within a block time-series cross-validation using k-NNR($k = 5$). We use the autoregressive order value of the model ($p = 2$) to determine the number of observations used as blocks between each fold. Recall that in a time-series cross-validation, each succeeding fold is a superset of the fold prior to it. As such, the final fold is the largest sample and hence contains the most information. Given this, for each signal we take its IC to be the predictive R^2 value measure over the final fold. The IC value used in the numerator of Equation 1 is simply the average final-fold IC of all the signals being considered. IC variance is defined as the cross-sectional variance of the four signals' IC values in the fifth fold.[14]

In Exhibit 2, we show the derivation of IR values for our hypothetical investment process. We first show the individual signal ICs in each of the five folds used in the block time-series cross-validated regression described in the previous section of the article. Next, we show the mean IC value and IC variance across the set of folds, followed by the cluster assignment for each signal. As the exhibit shows, the Mkt-Rf factor belongs to one cluster, while the other signals belong to a second cluster. Although the signals assigned to cluster 2 may be fundamentally different in terms of the market phenomena they are designed to measure, from the standpoint of their predictive properties in relation to the US manufacturing sector, they appear to be similar. Finally, we show the regime-specific IR for the set of signals using the procedure described in the preceding paragraph. The results show that the set of signals under consideration shows some predictive utility (positive IR) in the low growth regime but none in the high growth regime (though better than the naïve signal taken alone). In terms of individual factors, only Mkt-Rf exhibits consistently high levels of predictive efficacy across regimes. That said, the SMB and HML factors show themselves to be more predictively informative relative to the naïve signal.[15]

[14] Although the method used to calculate the IC and IC variance in the article has its merits, we realize that it is not the only way to arrive at inputs to the formula in Equation 1. For example, we could have taken the IC for each signal to be the average R^2 value across the five folds. The IC variance for each signal could likewise have been calculated across five folds, with the average IC variance of the four signals serving as the model IC variance used in the denominator in Equation 1.

[15] Note that the example presented in this article is for illustrative purposes only and not intended to be a comprehensive and final analysis of the predictive utility of the factors considered.

EXHIBIT 2

Derivation of Predictive Information Ratios for Fama–French and Naïve Signals (monthly data, August 1975–September 2019)

Regime	Mkt-Rf	SMB	HML	Naïve
High Growth				
IC Fold 1	0.827	–0.259	–0.808	–1.394
IC Fold 2	0.633	–0.467	–0.397	–0.392
IC Fold 3	0.669	–0.024	–0.117	–0.184
IC Fold 4	0.820	–0.151	–0.319	–0.117
IC Fold 5	0.824	–0.163	–0.270	–0.548
Mean IC	0.754	–0.213	–0.382	–0.527
IC var	0.007	0.022	0.054	0.211
Cluster	1	2	2	2
IR	–0.042			
Training Sample Observations	360			
Testing Sample Observations	72			
Low Growth				
IC Fold 1	0.827	–0.005	–0.047	–0.398
IC Fold 2	0.725	–0.060	–0.242	–0.307
IC Fold 3	0.878	0.125	0.088	–0.209
IC Fold 4	0.893	–0.737	0.024	–0.221
IC Fold 5	0.766	0.026	–0.453	–0.202
Mean IC	0.818	–0.130	–0.126	–0.268
IC var	0.004	0.096	0.039	0.006
Cluster	1	2	2	2
IR	0.261			
Training Sample Observations	82			
Testing Sample Observations	16			

CONCLUSION

Investment strategies are, at their core, aggregations of individual investment signals that form an alpha engine, much in the same way that the individual parts of a car engine work together to drive its performance on the road. Moreover, just like a car engine, the maintenance and improvement of an alpha engine requires diagnostic tools to adequately assess its performance in different environments, identify redundancies, and suggest areas for improvement. Consequently, model validation plays a central role in developing any investment process.

Although not originally intended as such, the formal framework known as The Fundamental Law of Active Management provides the conceptual building blocks for developing a model validation framework that is consistent with investment practice. To that end, this article introduces modular machine learning, an approach to model validation that is driven by the interdependence of a set of individually operating algorithms, each applied to a specific component of the model validation process. The first module applies a traditional econometric model to a data sample to generate a set of economically interpretable subsamples that function as the individual folds; we use these folds in block time-series cross-validated regressions in the second module to derive values for the information coefficient and the information coefficient variance for each investment signal under consideration. This information is then used as input to our third module, which uses cluster analysis to determine the level of breadth that exists within our set of investment signals. Breadth is defined as the number of clusters that result from our clustering operation, affinity propagation. Once the operations in each of our modules are complete, we use the output values for the information coefficient, information coefficient variance, and breadth to derive an information ratio for our signal set, which we interpret as an overall measure of investment model strength. Though modular machine learning is motivated by practical considerations, it also serves as an example of the potential to fruitfully combine econometrics and data science to produce synergies that can be positively additive to portfolio management.

REFERENCES

Agrawal, R., T. Imieliński, and A. Swami. "Mining Association Rules Between Sets of Items in Large Databases." *Proceedings of the 1993 ACM–SIGMOD'93*, pp. 207–216. New York: ACM Press, 1993.

Bergmeir, C., and J.M. Benítez. 2012. "On the Use of Cross-Validation for Time Series Predictor Evaluation." *Information Sciences* 191: 192–213.

Bergmeir, C., R. J. Hyndman, and B. Koo. 2018. "A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction." *Computational Statistics & Data Analysis* 120: 70–83.

Binmore, K. *Rational Decisions*, Princeton, NJ: Princeton University Press, 2009.

Boland, L. A. "Reading and Misreading Friedman's 1953 Methodology Essay." In *Milton Friedman: Contributions to Economics and Public Policy*, edited by R. Cord and D. Hammond, pp. 541–560. Oxford: Oxford University Press, 2016.

Breiman, L. 2001. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." *Statistical Science* 16 (3): 199–231.

Buckle, D. 2004. "How to Calculate Breadth: An Evolution of the Fundamental Law of Active Portfolio Management." *Journal of Asset Management* 4: 393–405.

Burman, P., E. Chow, and D. Nolan. 1994. "A Cross-Validatory Method for Dependent Data." *Biometrika* 81 (2): 351–358.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): 1–68.

Clarke, R., H. de Silva, and S. Thorley. 2002. "Portfolio Constraints and the Fundamental Law of Active Management." *Financial Analysts Journal* 58 (5): 48–66.

——. 2006. "The Fundamental Law of Active Portfolio Management." *Journal of Investment Management* 4 (3): 54–72.

Cover, T., and P. Hart. 1967. "Nearest Neighbor Pattern Classification." *IEEE Transactions on Information Theory* 13 (1): 21–27.

Ding, Z., and R. D. Martin. 2017. "The Fundamental Law of Active Management: Redux." *Journal of Empirical Finance* 43: 91–114.

Fama, E., and K. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

Filardo, A. J. 1994. "Business-Cycle Phases and Their Transitional Dynamics." *Journal of Business & Economic Statistics* 12 (3): 299–308.

Fix, E., and J. L. Hodges. "Discriminatory Analysis, Nonparametric Discrimination: Consistency properties." Technical report, USAF School of Aviation Medicine, 1951.

Friedman, M. "The Methodology of Positive Economics." In *Essays in Positive Economics*, pp. 3–43. Chicago: University of Chicago Press, 1953.

Frey, B. J., and D. Dueck. 2007. "Clustering by Passing Messages Between Data Points." *Science* 315 (5814): 972–976.

Ghazi, B., R. Panigrahy, and J. R. Wang. "Recursive Sketches for Modular Deep Learning." 36th Annual International Conference on Machine Learning, 2019.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37 (3): 424–438.

Grinold, R. C. 1989. "The Fundamental Law of Active Management." *The Journal of Portfolio Management* 15 (3): 30–37.

——. 1994. "Alpha is Volatility Times IC Times Score, or Real Alphas Don't Get Eaten." *The Journal of Portfolio Management* 20 (4): 9–16.

Grinold, R. C., and R. Kahn. *Active Portfolio Management,* 2nd ed. New York, NY: McGraw-Hill, 2000.

Hamilton, J. D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57 (2): 357–384.

Kahn, R. 1997. "Seven Quantitative Insights into Active Management Part 3: The Fundamental Law of Active Management," *BARRA Newsletter*, Winter.

Menchero, J. 2017. "Rethinking the Fundamental Law of Active Management." *Journal of Investment Management* 15 (2): 92–107.

Mézard, M. 2007. "Where are the Exemplars?" *Science* 315 (5814): 949–951.

Qian, E., and R. Hua. 2004. "Active Risk and Information Ratio." *Journal of Investment Management* 2 (3): 20–34.

Racine, J. 2000. "Consistent Cross-Validatory Model-Selection for Dependent Data: hv-Block Cross-Validation." *Journal of Econometrics* 99 (1): 39–61.

Roberts, D. R., V. Bahn, M. S. Boyce, S. Ciuti, C. F. Dormann, J. Elith, G. Guillera-Arroita, F. Hartig, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, and B. A. Wintle. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40 (8): 913–929.

Simonian, J., and F. Fabozzi. 2019. "Triumph of the Empiricists: The Birth of Financial Data Science." *The Journal of Financial Data Science* 1 (1): 10–13.

Simonian, J., and C. Wu. 2019. "Minsky vs. Machine: New Foundations for Quant-Macro Investing." *The Journal of Financial Data Science* 1 (2): 94–110.

Simonian, J., C. Wu, D. Itano, and V. Narayanam. 2019. "A Machine Learning Approach to Risk Factors: A Case Study Using the Fama–French–Carhart Model." *The Journal of Financial Data Science* 1 (1): 32–44.

Ye, J. 2008. "How Variation in Signal Quality Affects Performance." *Financial Analysts Journal* 64 (4): 48–61.

Zhou, G. 2008. "On the Fundamental Law of Active Portfolio Management: What Happens If Our Estimates Are Wrong?" *The Journal of Portfolio Management* 34 (4): 26–33.

Disclaimer
This article is for informational purposes only and should not be construed as legal, tax, investment, financial, or other advice. Nothing contained in the article constitutes a solicitation, recommendation, endorsement, or offer by the author to buy or sell any securities or other financial instruments.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Triumph of the Empiricists: *The Birth of Financial Data Science*

Joseph Simonian and Frank J. Fabozzi
The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/10

ABSTRACT: *The authors situate financial data science within the broader history of econometrics and argue that its ascendance marks a reorientation of the field toward a more empirical and pragmatic stance. They also argue that owing to the unique nature of financial information, financial data science should be considered a field in its own right and not just an application of data science methods to finance.*

Minsky vs. Machine: *New Foundations for Quant-Macro Investing*

Joseph Simonian and Chenwei Wu
The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/2/94

ABSTRACT: *Systematic macro investors' use of the regime-switching models that have been developed in academia over the last several decades is infrequent at best and, when used, generally tangential to their core investment process. The roots of this less-than-enthusiastic uptake can be found in two familiar sources: models that possess an overly complex formal structure and poor predictive ability. As a remedy to the current state of affairs, the authors present a new foundation for regime-based investing, one based on spectral clustering, a graph theoretic approach to classifying data. Drawing inspiration from the work of Hyman Minsky and John Geanakoplos, the authors present a macro framework that uses measures of growth, inflation, and leverage to define regimes and drive portfolio decisions. To the latter end, the authors show how the framework can be used to build portfolios using information about regimes as defined, to outperform a no-information equal-weight portfolio both out-of-sample and in bootstrapped and cross-validated simulations. The authors thus show that spectral clustering can provide both an elegant mathematical description of the leverage cycle and a robust foundation for quant-macro investing.*

A Machine Learning Approach to Risk Factors: *A Case Study Using the Fama–French–Carhart Model*

Joseph Simonian, Chenwei Wu, Daniel Itano, and Vyshaal Narayanam
The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/32

ABSTRACT: *Factor models are by now ubiquitous in finance and form an integral part of investment practice. The most common models in the investment industry are linear, a development that is no doubt the result of their familiarity and relative simplicity. Linear models, however, often fail to capture important information regarding asset behavior. To address the latter shortcoming, the authors show how to use random forests, a machine learning algorithm, to produce factor frameworks that improve upon more traditional models in terms of their ability to account for nonlinearities and interaction effects among variables, as well as their higher explanatory power. The authors also demonstrate, by means of a simple example, how combining the random forest algorithm with another machine learning framework known as association rule learning can produce viable trading strategies. Machine learning methods thus show themselves to be effective tools for both ex post risk decomposition and ex ante investment decision-making.*

Interactions and Interconnectedness Shape Financial Market Research

OTTO LOISTL AND GUEORGUI S. KONSTANTINOV

OTTO LOISTL
is professor emeritus of finance and chair of the Institute of Investment Banking and Catallactics at Vienna University of Economics in Vienna, Austria.
otto.loistl@wu.ac.at

GUEORGUI S. KONSTANTINOV
is senior portfolio manager in Fixed-Income & Currencies at LBBW Asset Management in Stuttgart, Germany.
gueorgui.konstantinov@lbbw-am.de

KEY FINDINGS

- Market performance at a real micro level becomes a key issue for modern asset management as order book analysis becomes vital, because transactions mean interactions. Big data, computer power, and machine learning allows analysis of this organized complexity.
- There are two approaches to modelling market performance: "from orders to prices" by doubly stochastic Markov processes and by measuring the interconnectedness of prices and markets via networks; these approaches can be complementarily used in modern research.
- Network analysis for factors and assets and centrality-based solutions in portfolio management are advantageous for capturing nonlinearity in data. Asset allocation decisions and risk management are examples to consider in machine learning.

ABSTRACT: *In this article, the authors investigate two fields that might be relevant for financial data sciences. The first issue covers the entire production chain "from orders to prices" by realistically modeling stock exchanges' microstructure (e.g., NASDAQ and Xetra). Specifically, the authors show how data-driven research can model decisions to place orders and to generate prices by matching orders accordingly. The other issue is price interconnectedness at markets by networks. The authors show that interactions shape a market's performance. Emergence comprises the interactions at markets; as such, the collective may not be equal to the sum of individual activities. As a consequence, the assumption that markets are in equilibrium and that arbitrage opportunities do not exist can be replaced by more realistic working hypotheses. The authors show with the two examples that market participants interact, learn, and trade. These individual interactions can be described as organized complexity. Whereas calculus may not support explicit modeling of interactions, the age of big data permits their modeling and application of innovative concepts, such as network solutions for asset allocation, which can be modeled using machine learning. This article illustrates that assertion with concrete examples.*

TOPICS: ***Equity portfolio management, statistical methods, simulations, big data/machine learning****

Huge data and strong computer power enable successful portfolio management. In fact, the continuous data flow permits investigations in finance we could only dream of 30 years ago. The assumptions that markets are in equilibrium and that arbitrage opportunities do not exist will be replaced by more realistic working hypotheses proposed

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

and tested by data-driven science. However, how these two market forces may be combined—and whether hitherto applied concepts are still useful in the face of new possibilities—remains an open question.

In this respect, machine learning is a promising tool. Increasing computer power may be unrivalled in its success in recognizing a learned pattern in unknown data, but how can untrained computers exploit unknown data without knowing the true objectives of portfolio management?

In its infancy, the data science paradigm's tools may not be as clear-cut and streamlined as the old one's have been. The proponents of the old paradigm may have difficulties getting acquainted with new point(s) of view. Defense of the old procedures is also motivated by fear of losing a competitive edge. No doubt machine learning will provide competitive advantages if the technique is applied appropriately, and we propose solutions toward such a goal. In our opinion, long-term and short-term views have to be differentiated with regard to the data needed. For short-term management, high-frequency trading, for example, investigation of order book analysis is an appropriate solution. Long-term investment decisions may rely on implementing data analysis into fundamental data—a *quantamental* investing approach. To express these ideas, we turn to a short overview of the development of the new financial research paradigm.

In this article, we investigate two fields that might be relevant for financial data sciences. The first covers the transition "from orders to prices" by realistically modeling stock exchanges' microstructure (e.g., NASDAQ and Xetra). Specifically, we show how data-driven research can model decisions to place orders and to generate prices by matching orders accordingly and predicting market performance. The other field involves price interconnectedness in markets by networks. More precisely, we show how factors and assets are connected in the market.

INTERACTIONS AND EMERGENCE

Organized Complexity and Emergence

A look at the history may provide some useful information. Weaver (1948) discerned "organized simplicity," "disorganized complexity," and "organized complexity" from one another and stressed the vital importance of "organized complexity" for both social and natural sciences in the 21st century. Notably, Hayek (1974) referred to Weaver's (1948) analysis in his Nobel Lecture when he articulated natural scientists' "pretense of knowledge" in tackling economic problems correctly. Hayek (1983, p. 251) reaffirmed that view nearly ten years later: "They are not mass phenomena of the kind to which statistical theory is applicable." Notably, Klir (2006) visualized this reasoning, which we have expanded to reflect the tendencies in data science and show in Exhibit 1.

EXHIBIT 1
The Complexity–Randomness–Relationship

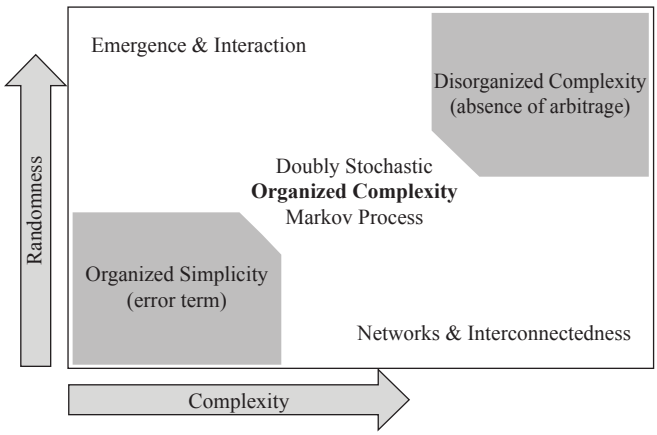


Exhibit 1 reflects "randomness" and "complexity." The problems of organized simplicity present randomness as an error term that diminishes in cases of optimization. Alternatively, the problems of disorganized complexity deal with the hitherto applied probability theory, based on linearity and referring to the appropriate measure theory, reflecting the assumption that no arbitrage opportunities exist. Sometimes these models come along with the assumption of equilibrium. However, how both assumptions fit together is an open question. Furthermore, Hayek (1974) underlined that the hitherto prevailing application of statistical criteria such as mean, variance, covariance, and skewness are appropriate to describe mass data problems located in the field of disorganized complexity. The analysis of economic problems located in the field of organized complexity, however, requires different analytical tools.

Weaver's (1948) prophesy that individuals and their specific preferences dominate research has become the focus of real problem solving: Big data powered the recording of individual behavior and the collection of

data about it. For example, data collection by smart phones, apps, and the Internet of things extends the dataset available for analysis. In fact, the analysis of the bulk of alternative data enables the transition from individual data to collective data. As far as we can see, these procedures rely on statistical correlations in the broad sense, without a theoretical foundation and without causal structures. In times of big data and big computers, this is a pragmatic and efficient method of financial research. We discuss this topic under the heading of interconnectedness. It may work accordingly, as long as there are no structural changes and the correlations remain reliable. However, a theoretical modeling of human behavioral attitudes might be more reliable as new data arrive. Therefore, we will present first an approach based on both analytically founded behavior and on factual structures of market data. In our framework, the sum of the collective may differ from the pure sum of individuals' actions. Interactions among individuals are essential. This is emergence.

Markets Mediate Micro and Macro Connections

The innovative issue, emergence, promoted by big data and big computer, may revolutionize economics to fulfill the vital postulate "that all explanations must run in terms of actions and interactions of individuals" (Arrow 1994, p. 1). In fact, Haken (1978, 1983), pioneered the issue of emergence in natural science by introducing "coupling parameters," which connect micro and macro variables and thus explain several phenomena in physics. Unfortunately, the search for coupling parameters according physics' philosophy failed in economics. Scientists ignored a market's coupling function to connect micro activities into macro variables.

In our view, both big data and emergence will revolutionize the modeling of the capital market, thus providing the missing link to connect micro and macro level consistently. We are confident that the true market activities at the micro level will be modeled in the future. To achieve this, theory has to abandon the assumption that markets are in equilibrium. Under such an assumption, it is neither necessary nor possible to investigate individuals' actions, their motivations, and, last but not least, their interactions (i.e., the action individuals take when observing what other market participants are doing). The whole market structure is not the sum of isolated variables.

The new technology era, in our view, forces finance and economics to investigate the vital issue of markets and not disregard it by abstract assumption. In this respect, two issues deserve attention. The first refers to investigating equidistant price time series instead of the motivations to trade. The second reflects the production chain and, in particular, how the transition from values to prices works. We propose a new way of thinking and provide a highlight in the investment management approach.

Loistl and Landes (1989) and their working group proposed a combination of Markov process and McFadden's (2001) *discrete choice* to model th mediation of interactions by markets.[1] The authors' objective was to model the actions of stock market participants as precisely as possible. The main idea is not to rely on simple time series modeled as a stochastic process. Instead, the authors focused on evaluating the Frankfurt Stock Exchange's traders' decision-making process regarding placing, accepting, and withdrawing orders. Combining Markov processes' transition rates with the decision probabilities provided by discrete choice successfully solved real trading possibilities (see Loistl and Landes 1989). Specifically, merging exponentially distributed Markov transition time with discrete choice probabilities—that a given action will be taken—enables modeling of trader interactions and learning.[2] It is a true representation of a trader's behavior.

Formally, the entire market performance is given by a master equation comprising the entire set of transition rates. Later, Huetl, Loistl, and Prix (2006) named this approach the *doubly stochastic Markov process* (DSMP) because it combines both Markov processes and a decision-making process operating according to discrete choice. The master equations can be handled numerically by the current capabilities of data processing. Currently, data science methods allow efficient implementation. Moreover, industrywide tests would also answer the question of whether theory is superfluous and if computers may overtake specific tasks. Einstein, according to Heisenberg (1969), once observed that it is always the theory that tells us what we have to measure. This aphorism is in line with the recent call by Arnott,

[1] Spitzer (1970) studied interaction modeling by Markov processes.

[2] Breiman (1969) showed that Markov transition time is exponentially distributed.

Harvey, and Markowitz (2019) for economic reasoning in using machine learning algorithms.

New implementation possibilities arise that, in our view, deserve attention. Loistl, Schossmann, and Vetter (2001) and Loistl et al. (2002) investigated transaction costs at Xetra and NASDAQ, simulating the market's activities under realistic trading circumstances. The model is appropriate to evaluate transaction costs at different trading places. Loistl et al. (2002) found that NASDAQ exhibits minimal transaction costs for mid-size trades, whereas transaction costs for small and block trades are nearly 100% and 50% higher.[3] Huber et al. (2004) noted that a logit model implies independence from irrelevant alternatives. The model's basic idea is also compatible with a probit model. Because logit and probit models find wide application in modern data-driven finance, we argue that the model can be implemented within an investment process and machine learning algorithm.

Key economic behavioral actions such as order placing and pricing activities have to be investigated. However, scholars may not be familiar yet with modeling interactions explicitly. In the absence of mathematical tools to model the connection of micro and macro variables consistently and efficiently in the last century, organized complexity will dominate both social and natural sciences in the first half of the 21st century. This applies to the discrete and real activities of market participants and their explicit interactions. Hayek's (1974) Nobel Prize lecture, "Pretense of Knowledge," was an early precursor of critiques of the Newtonian view of economics, which López de Prado and Fabozzi (2017) recently discussed. What we need are reliable theories.

RELIABLE THEORIES REQUIRE APPROPRIATELY STRUCTURED DATA

Science Moves from Reductionism to Emergence

Theories' empirical relevance was conditioned by the availability of scarce data. Tested with Newtonian methods, these theories are not appropriate to use the capabilities of the computers now available with regard to data exploration or numerical data processing. The old theories concentrated on, for example, integrals or differential equations that are analytically solvable. As López de Prado (2018a) pointed out, differential equations or integrals that can be solved analytically may not explain the reality sufficiently exactly in either natural or social sciences. Analytically solvable differential equations would not have guided astronauts correctly to the moon. This problem reflects the deficits of economics and economists raised by McFadden (1999): "Economists know the price of everything and the value of nothing." We need appropriate mathematical methods to build models that can exploit the ample data now available to investigate the new science's questions accordingly. The new science, both in natural and social sciences, may be termed emergence:

> Science has moved from an Age of Reductionism to an Age of Emergence, a time when the search for the ultimate causes of things shifts from the behavior of parts to the behavior of the collective.… The transition to the Age of Emergence brings to an End the myth of the absolute power of mathematics (Laughlin 2005, p. 208)

In this line of thought, Pines (2014) stressed that emergence is a unifying theme for 21st century science. Therefore, the modeling of interactions becomes vital to combine micro activities and macro variables consistently.

The Cost Effectiveness of Data

How big are the efforts to restructure data such that they are appropriate to solve practical problems? To put it differently: Is the data's price justified by the better performance achieved by better portfolio management? Similar questions were raised several years ago in the case of enterprise evaluations for initial public offerings (IPOs), for example. Because past value creation and price mechanisms might be questionable, data analysis now creates more traction on the true value. Most recently, Monk, Prins, and Rook (2019) discussed the importance of data in the investment management industry. The authors stressed the competitive advantage of big investment funds when exploiting data exclusively. There is a positive relationship between excludability and the value of data.

Because the value of alternative data is not always clear-cut and such data are costly, it is appropriate to look at publicly available data. To our knowledge, there exists no synopsis of all the relevant fundamental

[3] The simulation scenario was based on NASDAQ rules and regulations that were valid at that time. NASDAQ's management was aware of these circumstances: It checked the rules and regulations and kept NASDAQ's SOES but proposed to introduce Super Montage. See Loistl et al. (2002, pp 221).

data—at least not yet. Databases differ with regard to the data offered, their reliability, and last but not least the cost and accessibility. We discern two basic aspects: data referring to a company's fundamentals and data referring to the stock market performance. The attributed value will be obvious in case of insider information, which, according to its definition, has an impact on the price in the case they are made public.

From Fundamental Data to Quantamental Investing

Fundamental data may reflect a company's earnings power. Annual reports provide the most important information. Companies listed and engaged in interstate trade are required to deliver an annual report to the US Securities and Exchange Commission's (SEC) EDGAR (Electronic Data Gathering, Analysis, and Retrieval system) database. The challenging task of comparing different companies' reports has changed because the Extended Business Reporting Language (XBRL) fulfills the reporting duties. In fact, XBRL has been a common initiative of leading IT companies such as IBM and Microsoft and leading accounting firms. The data structure has been unified so that reports of different companies became comparable without major adjustment because of a unique taxonomy.

Tesfatsion (2007) was one of the first to articulate this necessity and highlighted the advantages of a taxonomy comprising economic theory. The taxonomy of annual reporting figures matured. The SEC, as the watchdog for annual reporting duties, took care of XBRL, namely to fulfill EDGAR reporting duties. Other institutions all over the world followed. As Golden (2019) put it: "XBRL has become an important gateway for the use of structured data that benefits investors and out capital markets." The key success of the taxonomy's development and maintenance responsibilities, achieved by the Financial Accounting Foundation (FAF) and the Financial Accounting Standards Board (FASB), is in our view a major step in the development of data-driven investing. It is the cornerstone in the implementation of the *quantamental investing*—integrating fundamental and data analysis.[4] Moreover, the application and implementation of fundamental factors in a data-driven investment process remains a possibility. With regard to accessibility and liability, the fundamental data provided by means of XBRL may have a competitive edge.

Another issue is the multifactor models promoted by Fama and French (1992, 1993, 2015) and Carhart (1997). The multifactor linear models are a mixture of fundamental and stock market data (Konstantinov and Rusev 2020). As factor investing has gained importance in recent years, machine learning and data science approaches are extensively used to construct portfolios (see Simonian et al. 2019; Konstantinov, Chorus, and Rebmann 2020). The applications of financial data science in investment management are widening. Jacobson, Lee, and Ma (2019) argued that investors are increasingly interested in the environmental, social, and corporate governance dimensions of their investments, which broadens data applicability. How can investment professionals use the data in practice? We provide examples, starting with order book analysis.

From Transaction Data to Vivid Market Performance: A Practitioner Example

In the past, capital market research centered on the investigation of price time series transformed into return figures. The statistical procedures assume mostly that the single price/return quotations are given as a straight time series and the individual figures are quoted equidistant. Research now focuses on market performance as such. Jackson (2013) and Roth (2015, 2018) provide a comprehensive description of different market functionalities. Price discovery is a major goal in addition to the goods transactions from sellers to buyers. To investigate the marketplaces and their microstructure accordingly, the data necessary have to be as detailed as possible. Good data enable good research—or the other way around (garbage in, garbage out). To be fair to the researchers: Approximate data only deliver approximate results. Therefore, the recording of individual orders is vital for correct performance.

How can investment professionals use the proposed analysis? In the following, we provide an example for data analysis to give an impression of how complete data might look after reordering and restructuring of the raw data. The data presented contain all entries in the Frankfurt Stock Exchange order book occurring on the six trading days from Monday, January 5, to

[4] Kothari (2019) mentioned many more electronic formats aside from XBRL, such as XML, FIX, FpML, and Inline XBRL. Standardization and taxonomy is an ongoing process, and it is beyond our scope to discuss XBRL or InXBRL further. For details see FASB (2019).

Exhibit 2

Frequency of Different Events in the Database

Event Type	Event Code	Absolute Frequency
Order Entry	1	2,284,628
Order Modification	2	36,165
Order Cancelation	3	1,626,896
Order Filled Completely	4	675,232
Order Partially Filled	5	373,760
Order Deleted Automatically	6	16,597
Technical Entry	101	461
Technical Cancelation	103	461

Source: Data are adapted from Prix, Loistl, and Huetl (2007).

Monday, January 12 2006.[5] This time span has been chosen to allow comprehensive study of the weekend effect. As the data lay bare, there is not a weekend effect from Friday closing to Monday opening, but rather from Monday opening to Monday closing. This specific price movement is not detectable by observing only closing prices. All the investigated data comprise the complete record of all DAX entries during that time span. The data consist of 5,014,200 data points, recording all changes in the Xetra's electronic order book. The data entries consist of different registered event types. Each event type's code and frequency are listed in Exhibit 2. Prix, Loistl, and Huetl (2007) compared several samples of six-day trading during 2004–2007.[6]

Structuring enables us to check the effectiveness of the exchange's rules and regulations. Exhibit 3 illustrates the data after restructuring the raw data. It comprises orders, their limits, and the transactions completed during the time span under consideration. Exhibit 3 presents the events during the time span 10:30:08.35–10:30:12.56.

What steps are necessary to understand order book analysis? Without investigating the recorded orders, a reliable evaluation of the effectiveness of the exchange's price discovery is not possible. The data recording at other exchanges might be organized differently in detail. Prix, Loistl, and Huetl (2007) showed that it is possible to reorganize the data of a stock exchange such that every trading detail can be investigated. Nevertheless, bookkeeping requires the same data all over the world, which might be an advantage in implementing order book analysis across different markets. Specifically, it is possible to identify the time span of how long an order remains in the order book and to identify the causes of an order's extinction (see Prix, Loistl, and Huetl 2007, p. 727). The complete history of every trade—from placing orders to completing the transaction and recording the transaction's price—provides reliable information about the market performance. Kopp, Loistl, and Huetl (2008) and Kopp et al. (2008) showed that order placement reacts faster and more sensibly to stock-specific messages compared with price reactions. The micro situation of a market might be investigated for each stock. We provide an example in Exhibit 4: Extending the customary chart of orders, their limits, and volume into a cumulated presentation reveals further insights. We adopt the Paarsch and Hong (2006) approach to construct cumulated probabilities from discrete probability distributions.

Emergence of Cumulated Orders during Trading Time

Extending the graph of both cumulated orders by the time dimension reveals an even more informative picture (Exhibit 4). We term Exhibit 5 a bifurcated mountain because it has two slopes and a valley between. The valley's bottom between the slopes of cumulated bids and asks marks the spread between best bid and best offer. An observant reader can easily recognize that prices are generated by matching bids with asks, marked by a crossed circle. When a trade occurs, the spread shrinks to zero; either an ask reaches the bid slope or a bid reaches the ask slope. Trader-initiated or buyer-initiated trades will be clearly identified according to price time priority. What are the economic applications of this analysis?

A three-dimensional presentation of a market's performance conveys far more information than a pure (equidistant) time series of relative prices can provide. The next step is to generate a dynamic three-dimensional graph of a market's performance. Within an analytical process, the first step would be to generate orders by means of DSMP simulations. The second step would be to match orders according to the Xetra rules and regulations. Loistl (2011) and Huetl, Loistl, and Prix (2006) argued that by applying these steps, it is possible to predict the entire market's performance.

[5] We are much obliged to Frankfurt Stock Exchange for providing the data.

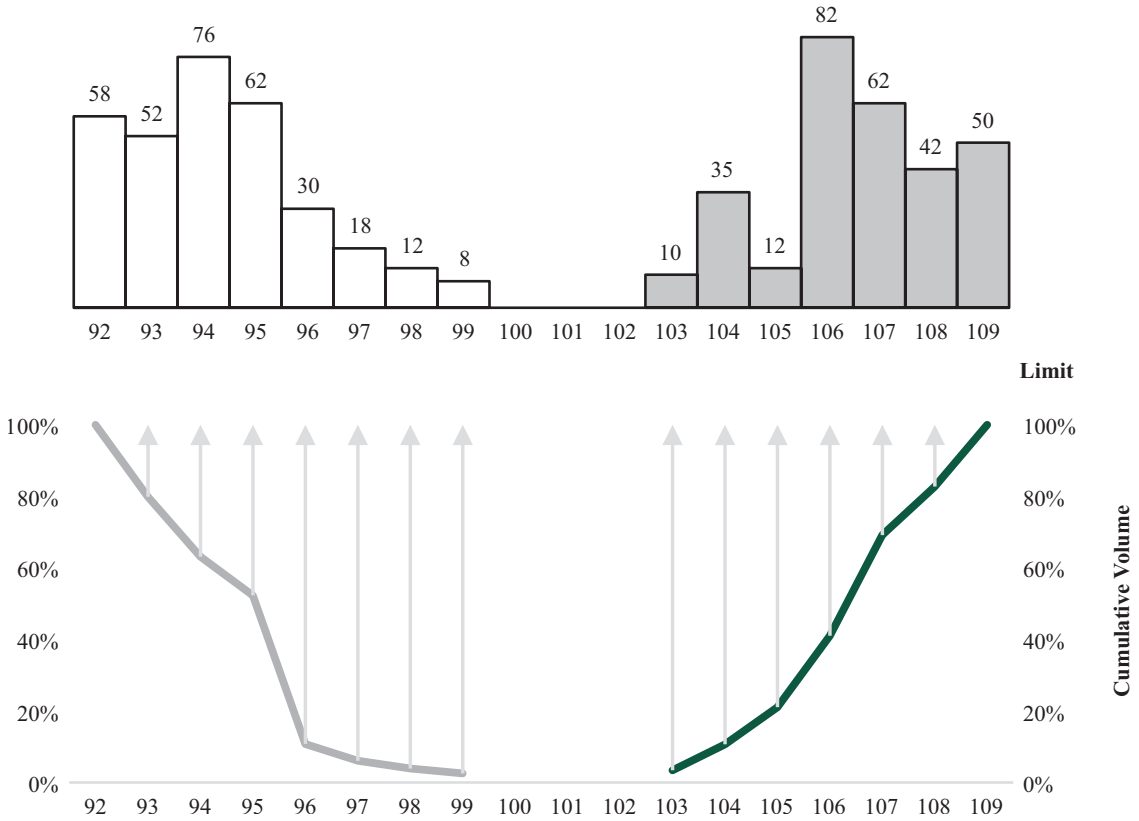
[6] The statistics of the four samples look quite similar (see Loistl 2011).

Exhibit 3
Bid, Ask, and Transactions

	Bid	Ask	Transaction
Time	**(size); Limit; Action; #**	**#; Action; Limit; (size)**	**#; Limit; (size); #**
10:30:08.35	(36); 97.53; Insertion; 7		
10:30:09.87		8; Insertion; 97.56; (600)	
10:30:11.89		9; Insertion; 97.55; (250)	
10:30:11.96	(400); 97.53; Insertion; 10		
10:30:12.56	(300); 97.52; Insertion; 21		
10:30:12.56		11; Accept; 97.53; (436)	11; 97.53; (36); 7
10:30:12.56			11; 07.53; (400); 10

Note: Data presented are from Deutsche Börse.
Source: Loistl (2011).

Exhibit 4
Cumulated Individual Orders

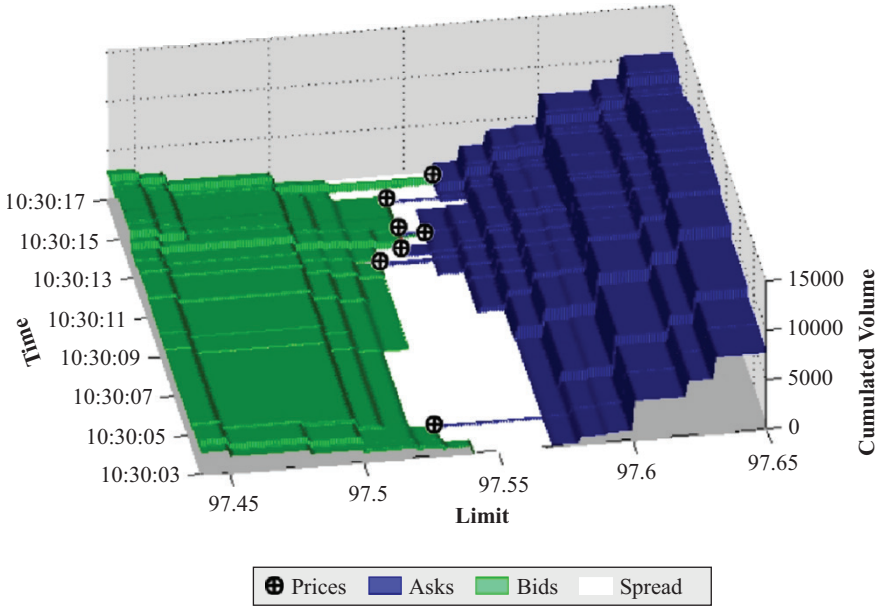


A Notion on Rules and Regulations

In this line of reasoning, we would like to point out several regulatory issues. Specifically, a reliable market surveillance system has to refer to orders and not merely to prices (Exhibits 4 and 5). Investigating orders and their limits enables market surveillance to hinder the development of what is sometimes called a *flash crash*. Looking at orders, such a danger is easy to detect. It is

EXHIBIT 5

Bifurcated Mountain—The Emergence of Cumulated Orders



then the duty of the market surveillance to propose a trading halt until the market calms. Therefore, an efficient market surveillance should have circuit breakers.[7]

NETWORKS AND CONNECTIVITY

From Newtonian Mathematics to Topology, Networks, and Entropy

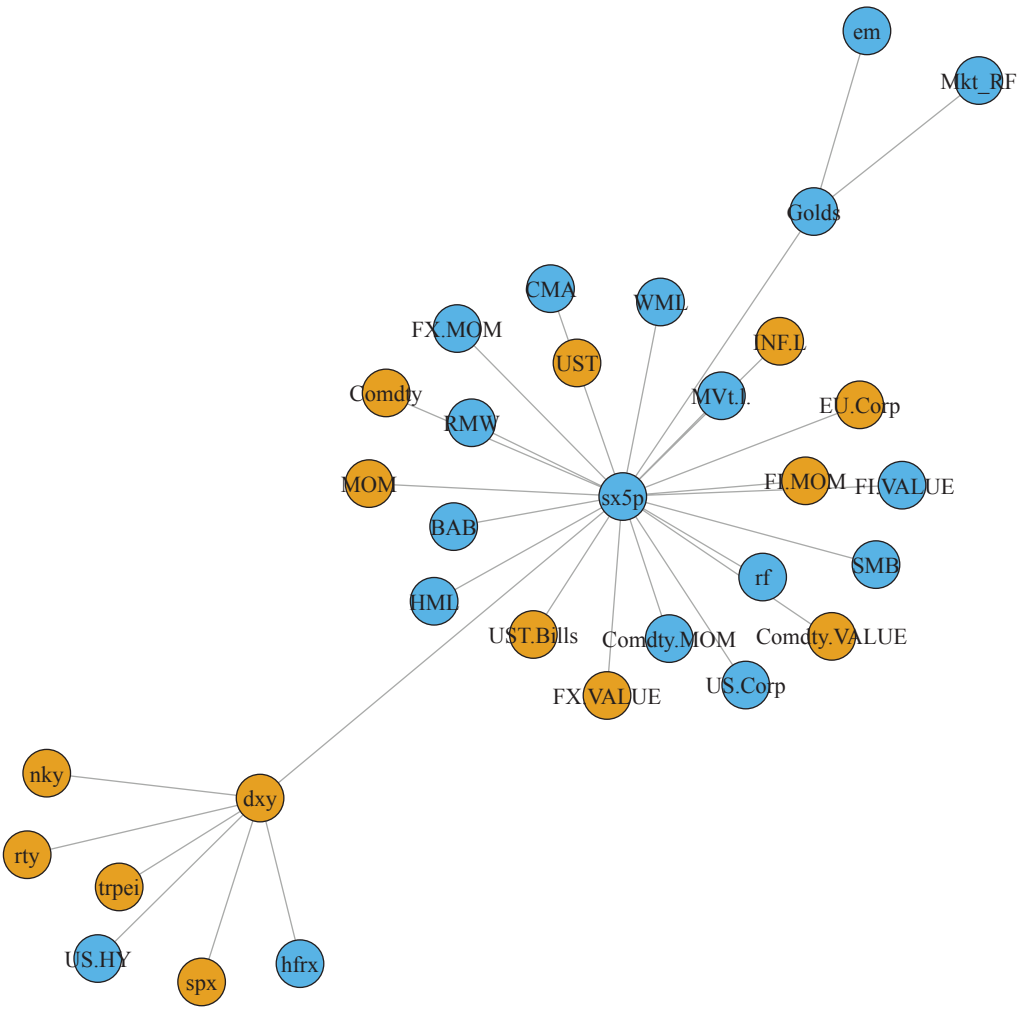
Nevertheless, there may be alternatives to DSMP to handle indications of interactions by price interconnectedness. Which interrelations are of particular interest? Asness, Moskowitz, and Pedersen (2013) argued that there is value and momentum in factor investing everywhere. Do we need calculus to prove that? In line with this view, Das, Kim, and Ostrov (2019) and Konstantinov, Chorus, and Rebmann (2020) stressed that factor and asset connectedness can be shown topologically in a network context. In fact, networks are one of the potential tools to comprehend that relation. Why? Topology is the science of connections. Thus, interrelations of entities are best represented with their mutual links in a network, which comprises nodes and edges. Using graph theory, analysts can show connectivity on both node and edge levels. Moreover, the connectivity of objects has properties, such as centrality scores and cluster affiliation, that help researchers to manage risks and portfolios by accounting for interrelations and dependencies. As shown in previous research in data science, machine learning techniques provide portfolio managers with different tools to asset and factor allocation. These techniques are indeed beyond mainstream calculus. Do we need perfect symmetries—normal distribution of returns, correlations, linearity, or markets in equilibrium—to build portfolios in the new millennium?

We believe we do not. López de Prado (2018a, 2018b) argued there exists no proof that the factor premiums are linear. Therefore, financial data analysis should account for nonlinearity. Financial data research needs tools that deal with interrelations and thus with organized complexity. For the moment, machine learning is far ahead of theoretical modeling, as argued by López de Prado (2018b). An example of interconnected assets and factors is shown in Exhibit 6. The observant reader can easily identify the two clusters and the connectedness between assets and factors. Specifically, currencies (DXY) and European equities (sx5p) connect the two major communities.

[7] See, for example, the rules and regulations of the Deutsche Börse in Frankfurt.

EXHIBIT 6

Clustered Correlation Network of Assets and Factors 2003–2017



Notes: The minimum spanning tree network has 32 nodes (both assets and factors) in the period from January 2003 to December 2017; we compute the distance metrics between two nodes from the correlation coefficients between them as given by the equation $d_{ij,t} = \sqrt{2(1-\rho_{ij})}$, *with* $i \neq j$. *We use the cluster spinglass algorithm, which is suitable for directed graphs with highest modularity score. The first cluster comprises following factors and assets: Euro Stoxx 50 (sx5p), emerging markets (em), US high yield bonds (US.HY), PE, Mkt_RF, SMB, RMW, CMA, WML, Gold (Golds), the liquidity premium (MVt.1), commodity momentum (Comdty.MOM), the book-to-market factor (HML), the betting-against-beta premium (BAB), the hedge fund Index (hfrx), the commodity momentum factor (Comdty.MOM), the US Corporate Bond Index (US.Corp), and the risk-free rate (rf). The second cluster comprises the Russell 2000 Index (rty), NIKKEI 225 Index (nky), US Dollar Index (dxy), S&P 500 Index (spx), commodity index (Comdty), momentum factor (MOM), US Treasury Index (UST), inflation-linked bond index (INF.L), European Investment Grade Corporate Bond Index (EU.Corp), US Treasury Bills Index (UST.Bills), fixed-income momentum factor (FI.MOM), currency value factor (FX.VALUE), commodity value factor (Comdty. VALUE), and private equity index (trpei). We use the "igraph" package in R software. Konstantinov, Chorus, and Rebmann (forthcoming) will explain the differences between directed and correlation-based networks.*

Sources: Prepared by the autors from Bloomberg Finance L.P., and AQR Capital Management, LLC, and Kenneth French's website data.

The diversification aspect of gold (Golds) can be easily seen as a separate node, connecting the Fama–French market (Mkt_rf) factor.

Why is network analysis important? López de Prado and Fabozzi (2017) highlighted the need for alternative methods in finance. Network analysis provides insightful information regarding factor-based connectedness, relationships, and how risk is transferred between network components. Both academic and practitioner research documents crowdedness in factor investing.

Consequently, risks are transferred within the web. Single entities, such as funds carrying risk factor or asset exposure, bear and may transfer risks to the whole market.

In fact, Arnott et al. (2019) stressed that factors are interconnected. Konstantinov, Chorus, and Rebmann (2020) showed, for example, a centrality-based network view of assets and factor allocation and how investors can build factor portfolios of interconnected components. Thus, researchers and portfolio managers using machine learning and big data can build models that incorporate asset and factor connectivity. Generating alpha means constantly using the constant information flow inherent to and driven by big data. Portfolio allocation, risk management, order book analysis and trading, regulatory technology, and so on are all steps from the epistemological, through data processing and classification to evaluation analysis, all of which fully reflect the issues pointed out by López de Prado (2018a) and are integral parts of modern asset management. Why? As pointed out by Grossmann and Stiglitz (1980), the markets cannot be efficient because new information is constantly arriving. Therefore, equilibrium can only be achieved for a very short moment until new information arrives and or the decision frame changes (see Monk, Prins, and Rook 2019). Objections must be made against game theory, even if it seems to tackle the economic core issue and investigate interactions. However, the basic scenario is determined by the requirement of an equilibrium that overrules interactions. Equilibrium scenarios are unrealistic, not only as new information constantly arrives but also as the game players try to influence the parameters and their arsenal of actions speeds up as in an arms race (see Schelling 1980). Therefore, using information flow and the dynamic market structure are keystones of big data and machine learning.

Integrated Data Flow and Continuous Decision Making

López de Prado (2018a) and Arnott, Harvey, and Markowitz (2019) showed why machine learning algorithm funds fail and identified the necessary steps in big data analysis. Economic theory plays a significant role. Therefore, measuring the interconnectedness between assets, factors, hedge funds, mutual funds, financial institutions, regulatory bodies, authorities, and the like remains a challenging and necessary issue, in our view. Both micro- and macro-prudential analysis would highlight the important cross-relations in a digitalized world. Kinlaw, Kritzman, and Turkington (2012) and Konstantinov and Rebmann (2019), among others, showed that determining systemic importance is crucial for the financial markets. Centrality-based solutions depart from mainstream calculus. The reliance on factor analysis is vital; however, in the era of big data, as centrality or importance scores are derived with appropriate mathematical tools that reveal the market structure, the view shifts from microscopic factor decomposition to the mesoscopic interaction-based structure embedded in the markets.

CONCLUDING REMARKS

Finance is moving away from perfect symmetries. We show this transition with two examples, which we argue might be successfully implemented in the industry. This is a shift from analytical calculus to topological discrete analysis. Python, the appropriate language for such a shift, offers a bright scope for analytical work. We do not doubt that defensible and defensive strategies will be in the major spotlight for investors. In this respect, big data, both investment and operational alpha, remain in the domain of data-driven investment managers. Both information flow and new techniques are suitable to deal with continuous market dynamics. This is organized complexity. This is order from chaos (see Simonian, López de Prado, and Fabozzi 2018). A new era in finance has finally arrived. A new era in finance will at least cover three issues: concentrating on actions and interactions of organized complexity, modelling the emergence of macro variables out of micro activities, and, thus, managing order out of chaos in the spirit of Simonian et al. (2018).

REFERENCES

Arnott, R., C. R. Harvey, V. Kalesnik, and J. Linnainmaa. 2019. "Alice's Adventures in Factorland: Three Blunders That Plague Factor Investing." *The Journal of Portfolio Management* 45 (4): 18–36.

Arnott, R., C. R. Harvey, and H. Markowitz. 2019. "A Backtesting Protocol in the Era of Machine Learning." *The Journal of Financial Data Science* 1 (1): 64–72.

Arrow, K. J. 1994. "Methodological Individualism and Social Knowledge." *American Economic Review* 84: 1–9.

Asness, C. S., T. Moskowitz, and L. H. Pedersen. 2013. "Value and Momentum Everywhere." *The Journal of Finance* 68 (3): 929–985.

Breiman, L. *Probability and Stochastic Processes: With a View toward Applications*. Boston, MA: Houghton Mifflin, 1969.

Carhart, M. M. 1997. "On Persistence in Mutual Fund Performance." *The Journal of Finance* 52 (1): 57–82.

Das, S. R., S. Kim, and D. Ostrov. 2019. "Dynamic Systemic Risk: Networks in Data Science." *The Journal of Financial Data Science* 1 (1): 141–158.

Fama, E. F., and K. R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–465.

——. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

——. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116: 1–22.

FASB Financial Accounting Standards Board. "XBRL US DQC Rules Taxonomy Technical Guide." US GAAP Financial Reporting Taxonomy, October 17, 2019.

Golden, R. G. "Remarks of the FASB Chairman Russell G. Golden." XBRL US Investor Forum: Driving Actionable Analytics, Baruch College, November 4, 2019.

Grossmann, S., and J. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70 (3): 393–408.

Haken, H. *Synergetics. An Introduction. Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology*, Berlin, Germany: Springer, 1978.

——. *Advanced Synergetics. Instability Hierarchies of Self-Organizing Systems and Devices*. Berlin, Germany: Springer, 1983.

Hayek, F. A. 1974. "The Pretense of Knowledge." Nobel Prize lecture December 11, 2019.

——. 1983. "The Keynes Centenary: The Austrian Critique." *The Economist*, June 11, 1983.

Heisenberg, W. 1969. "The Part and the Whole." Conversations in the Radius of Nuclear Physics, München, Germany [in German].

Huber, S., F. Leisch, O. Loistl, and A. Veverka. 2004. "Multidimensional Homogenous Markov Processes with Logit Transition Probabilities: A General Framework for Market Microstructure Modelling." Working paper, Vienna University of Economics and Business, 2004.

Huetl, M., O. Loistl, and J. Prix. "Quantitative-Behavioral Modelling of High-Frequency Market Data by Means of Doubly Stochastic Markov Process." Paper presented at the 13th International Conference on Forecasting Financial Markets, May 31–June 2, 2006.

Jacobson, B., W. Lee, and C. Ma. 2019. "The Alpha, Beta, and Sigma of ESG: Better Beta, Additional Alpha?" *The Journal of Portfolio Management* 45 (6): 6–15.

Jackson, M. O. 2013. "Economic Engineering and the Design of Matching Markets: The Contribution of Alvin E. Roth." *The Scandinavian Journal of Economics* 115 (3): 619–639.

Kinlaw, W., M. Kritzman, and D. Turkington. 2012. "Toward Determining Systemic Importance." *The Journal of Portfolio Management* 38 (4): 100–111.

Klir, G. *Uncertainty and Information: Foundations of Generalized Information Theory*. Hoboken, NJ: Wiley-Interscience, 2006.

Konstantinov, G., A. Chorus, and J. Rebmann. 2020. "A Network and Machine Learning Approach to Factor, Asset, and Blended Allocation." *The Journal of Portfolio Management* (forthcoming).

Konstantinov, G., and J. Rebmann. 2019. "From Risk Factors to Networks: A Case Study on Interconnectedness Using Currency Funds." *The Journal of Financial Data Science* 1 (3): 108–123.

Konstantinov, G., and M. Rusev. 2020. "The Bond-Equity Fund Relation Using the Fama-French-Carhart Factors: A Practical Network Approach." *The Journal of Financial Data Science* 2 (1): 24–44.

Kopp, E., M. Huetl, O. Loistl, and J. Prix. "Cross Sectional Liquidity Interactions from Intraday Perspectives." Paper presented at the Euro Working Group on Financial Modeling (EWGFM) meeting, 2008.

Kopp, E., O. Loistl, and M. Huetl. "The Influence of Real-Time Stock Specific Messages on Liquidity, Return and Trading Behavior," Working paper, Vienna University of Economics and Business, 2008.

Kothari, S. P. "Policy Challenges and Research Opportunities in the Era of Big Data." Big Data and High-Performance Computing for Financial Economics, National Bureau of Economic Research, Cambridge, MA, July 13, 2019.

Laughlin, R. *A Different Universe: Reinventing Physics from the Bottom Down*. New York, NY: Basic Books, 2005.

Loistl, O. "Agent-Based Modeling of Stock Exchanges by Means of Discrete Choice." Working paper, Vienna University of Economics and Business Administration, 2011.

Loistl, O., and T. Landes, ed. *The Dynamic Pricing of Financial Assets*. Hamburg, Germany: McGraw-Hill, 1989.

Loistl, O., B. Schossmann, and O. Vetter. 2001. "Xetra Efficiency Evaluation and NASDAQ Modelling by KapSyn." *European Journal of Operational Research* 125: 270–295.

Loistl, O., B. Schossmann, O. Vetter, and A. Veverka. 2002. "A Comparison of Transaction Costs on Xetra and on NASDAQ." *Quantitative Finance* 3: 199–216.

López de Prado, M. 2018a. "The 10 Reasons Most Machine Learning Funds Fail." *The Journal of Portfolio Management* 44 (5): 120–133.

——. *Advances in Financial Machine Learning*. Hoboken, NJ: Wiley, 2018b.

López de Prado, M., and F. J. Fabozzi. 2017. "Who Needs Newtonian Finance?" *The Journal of Portfolio Management* 44 (1): 1–3.

McFadden, D. 1999. "Rationality for Economists?" *Journal of Risk and Uncertainty* 19: 73–105.

——. 2001. "Economic Choices." *The American Economic Review* 91 (3): 351–378.

Monk, A., M. Prins, and D. Rook. 2019. "Rethinking Alternative Data in Institutional Investment." *The Journal of Financial Data Science* 1 (1): 14–31.

Paarsch, H. J., and H. Hong. *An Introduction to Structural Econometrics of Auction Data*. Cambridge, MA: MIT Press, 2006.

Pines, D. "Emergence: A Unifying Theme for the 21st Century Science." Santa Fe Institute, October 31, 2014.

Prix, J., O. Loistl, and M. Huetl. 2007. "Algorithmic Trading Patterns in Xetra Orders." *The European Journal of Finance* 13 (8): 717–739.

Roth, A. E. *Who Gets What—and Why: The New Economy of Matchmaking and Market Design*. New York, NY: Houghton Mifflin Harcourt, 2015.

——. 2018. "Marketplaces, Markets, and Market Design." *American Economic Review* 108 (7): 1609–1658.

Schelling, T. C. *The Strategy of Conflict*. Cambridge, MA: Harvard University, 1980.

Simonian, J., M. López de Prado, and F. J. Fabozzi. 2018. "Order from Chaos: How Data Science Is Revolutionizing Investment Practice." *The Journal of Portfolio Management* 45 (1): 1–4.

Simonian, J., C. Wu, D. Itano, and V. Narayanam. 2019. "A Machine Learning Approach to Risk Factors: A Case Study Using the Fama-French-Carhart Model." *The Journal of Financial Data Science* 1 (1): 32–44.

Spitzer, F. 1970. "Interaction of Markov Processes." *Advances in Mathematics* 5: 246–290.

Tesfatsion, L. 2007. "Agents Come to Bits: Towards a Constructive Comprehensive Taxonomy of Economic Entities." *Journal of Economic Behavior & Organization* 63 (2): 333–346.

Weaver, W. 1948. "Science and Complexity." *American Scientist* 36: 536–544.

Disclaimer

The opinions and views expressed in this article are those of the authors and may not necessarily represent the views of LBBW Asset Management or any of its affiliates or employees.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Alice's Adventures in Factorland: *Three Blunders That Plague Factor Investing*

Rob Arnott, Campbell R. Harvey, Vitali Kalesnik, and Juhani Linnainmaa

The Journal of Portfolio Management

https://jpm.pm-research.com/content/45/4/18

ABSTRACT: *Factor investing has failed to live up to its many promises. Its success is compromised by three problems that are often underappreciated by investors. First, many investors develop exaggerated expectations about factor performance as a result of data mining, crowding, unrealistic trading cost expectations, and other concerns. Second, for investors using naive risk management tools,*

factor returns can experience downside shocks far larger than would be expected. Finally, investors are often led to believe their factor portfolio is diversified. Diversification can vanish, however, in certain economic conditions when factor returns become much more correlated. Factor investing is a powerful tool, but understanding the risks involved is essential before adopting this investment framework.

A Backtesting Protocol in the Era of Machine Learning

Rob Arnott, Campbell R. Harvey, and Harry Markowitz

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/64

ABSTRACT: *Machine learning offers a set of powerful tools that holds considerable promise for investment management. As with most quantitative applications in finance, the danger of misapplying these techniques can lead to disappointment. One crucial limitation involves data availability. Many of machine learning's early successes originated in the physical and biological sciences, in which truly vast amounts of data are available. Machine learning applications often require far more data than are available in finance, which is of particular concern in longer-horizon investing. Hence, choosing the right applications before applying the tools is important. In addition, capital markets reflect the actions of people, who may be influenced by the actions of others and by the findings of past research. In many ways, the challenges that affect machine learning are merely a continuation of the long-standing issues researchers have always faced in quantitative finance. Although investors need to be cautious—indeed, more cautious than in past applications of quantitative methods—these new tools offer many potential applications in finance. In this article, the authors develop a research protocol that pertains both to the application of machine learning techniques and to quantitative finance in general.*

Dynamic Systemic Risk: *Networks in Data Science*

Sanjiv R. Das, Seoyoung Kim, and Daniel N. Ostrov

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/141

ABSTRACT: *In this article, the authors propose a theory-driven framework for monitoring system-wide risk by extending data science methods widely deployed in social networks. Their approach extends the one-firm Merton credit risk model to a generalized stochastic network-based framework across all financial institutions, comprising a novel approach to measuring systemic risk over time. The authors identify four desired properties for any systemic risk measure. They also develop measures for the risks created by each individual institution and a measure for risk created by each pairwise connection between institutions. Four specific implementation models are then explored, and brief empirical examples illustrate the ease of implementation of these four models and show general consistency among their results.*

A Holistic Approach to Financial Data Science: *Data, Technology, and Analytics*

TAMER KHRAISHA

TAMER KHRAISHA is a software and data developer at Alphacruncher Hungary Kft in Budapest, Hungary. **tamer.khraisha@alphacruncher.com**

KEY FINDINGS

- Financial data science is presented as an emerging field that complements classical financial econometrics.
- The financial data ecosystem is discussed in detail, illustrating the types and structures of financial data and their distinguishing features.
- A strategy of nine interrelated parts is proposed to efficiently manage a financial data science project. The proposed approach can be used by industry professionals and academic researchers.

ABSTRACT: *The scientific analysis of financial data, both for practical and theoretical purposes, has continuously been an active and evolving area. Traditionally, financial modeling and analysis have primarily relied on financial econometrics, which applies statistical methods to the problems of finance. Recently, advances in analytical and algorithmic methods, powerful technological solutions, and the expansion of the financial data ecosystem have created new opportunities and provided novel and robust ways to solve existing and new problems in finance. To expand the scope of financial econometrics and incorporate these novelties, the field of financial data science (FDS) has emerged. Crucially, FDS adds more interdisciplinary, methodological variety and technological components to the analysis process, thus generating the need for a holistic view in the management of an FDS project. In this article, the author presents a holistic and structured approach to efficiently and effectively manage an FDS project. The main contribution of the article is to provide a comprehensive, financial-domain–oriented approach to FDS management that can be used by researchers and practitioners to exploit the opportunities arising from the evolving finance landscape.*

TOPICS: ***Statistical methods, simulations, big data/machine learning****

Professionals and researchers in financial markets have always been among the early adopters of innovations in computer science, analytical methods, and data collection techniques. At present, nothing characterizes the evolution of the financial analysis landscape more than the employment of machine learning and data mining techniques; the increasing reliance on open source, cloud, and scalable technologies; and the use of large and alternative datasets. In this article, the author uses the term *financial data science* (FDS) to refer to this current trajectory in finance and defines it as follows:

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

> Financial data science is a distinct, interdisciplinary area of research and practice that combines tools and methods from financial economics, statistics, computer science, machine learning, and data mining to scientifically analyze and understand a wide variety of datasets to solve existing and new problems in finance.

Implicit in this definition are a few assumptions. To begin, the author follows Brooks et al. (2019) to claim that FDS is an interdisciplinary field, in that concepts, tools, and methods are borrowed from different and correlated areas of inquiry.[1] Interdisciplinarity should not mean that FDS is simply a mix of tools and methods from different disciplines; rather, FDS is to be treated as a field of its own with a strong emphasis on domain knowledge (Donoho 2017; Simonian and Fabozzi 2019). Finance has its unique set of problems and puzzles, and financial data possess particular characteristics that distinguish them from other types of data.

Financial analysis has extensively relied on the use of statistical methods to analyze financial data, mainly through financial econometrics (Tsay 2010). The substantial reliance on statistics in finance is due to the significant amount of uncertainty present in both financial data and financial theory and the practical impossibility of controlled experimentation in finance (Lo and MacKinlay 1990; Tsay 2010).

Theoretically speaking, the main goal of econometrics is to find causality relationships and uncover economic dynamics and mechanisms (Varian 2014a; Simonian and Fabozzi 2019). In econometrics, the most common tools of analysis are multivariate linear regression (Varian 2014b; López de Prado 2018), parameter estimation, and hypothesis testing (Mullainathan and Spiess 2017; Simonian and Fabozzi 2019). Financial econometrics has produced a wide variety of models that have been used by industry practitioners and regulators. However, recently, certain shortcomings of financial econometrics have come under criticism. For instance, several practices of multiple hypothesis testing and backtesting have produced biased and potentially false discoveries, mainly due to selection bias under multiple testing and the pressure to produce significant results (Kim and Ji 2015; Harvey, Liu, and Zhu 2016; Harvey 2017; López de Prado 2019). Additionally, financial econometrics relies on a limited number of models (e.g., regression models) that work reliably in certain contexts but fail to model complex and high-dimensional data (Varian 2014b; López de Prado 2018). Moreover, econometrics lacks the tools and methods to analyze alternative datasets such as social media streams, geolocation data, patents, news, and microdata on consumer behavior (Guida 2018; Monk, Prins, and Rook 2019).

This is where fields such as machine learning and data mining offer a whole different range of robust and innovative solutions that can complement financial econometrics.[2] Machine learning concerns itself with the development of algorithms and methods that can be trained on data, learn from experience, and adapt to environmental changes. In this regard, two particular differences between machine learning and financial econometrics are worth mentioning. First, while the main focus of financial econometrics is parameter estimation (i.e., understanding the data-generating process), machine learning emphasizes algorithmic out-of-sample prediction (Breiman 2001; Mullainathan and Spiess 2017; Van Liebergen 2017). Second, following the classification offered by Hey, Tansley, and Tolle (2009), financial econometrics falls within the empirical paradigm, which focuses on the verification of theoretical constructs and relies on assumptions, whereas machine learning falls within the data-intensive paradigm, whose main goal is recognizing patterns in data. This allows machine learning to have a wider range of applications, such as the automation of financial tasks and the creation of data products.[3]

A field related to machine learning is data mining. Data mining and machine learning share many concepts

[1] The interdisciplinarity of finance is not something entirely new: Areas such as computational finance or financial engineering are interdisciplinary in that they integrate numerical methods from mathematics and computer science (Lyuu 2001; Seydel and Seydel 2006; Omur 2008).

[2] FDS is not an alternative to financial econometrics. Rather, it expands the scope of financial econometrics to incorporate the novelties in the financial analysis landscape. In the end, most innovations are merely a combination of existing ideas/technologies that produce a new thing.

[3] Financial professionals have always created data products; for example, financial engineers use financial data to develop products such as derivatives. With machine learning, the novelty consists of those data-driven applications that derive their value from the data itself and generate more data as a result; examples of the latter are fraud detection systems and recommender systems.

and methods, with the slight difference being that the goal of data mining is more concerned with the automated discovery of knowledge and insights from large and complex databases (Fayyad, Piatetsky-Shapiro, and Smyth 1996). Financial analysis has extensively applied methods from data mining and machine learning, which have been proven to outperform more traditional tools at several tasks (Bose and Mahapatra 2001; Kovalerchuk and Vityaev 2005; Chen and Du 2009; Lahmiri and Bekiros 2019; Sirignano and Cont 2019).

Finally, a critical enabler and a primary driver of FDS is computer science. In the financial analysis landscape, computer science aspects have mainly concerned data storage and computation (Lyuu 2001; Seydel and Seydel 2006; Omur 2008). In this regard, the main novelty concerns the need for scalability. Typically, financial econometricians would have the necessary skills to store the data and code and run one or a few linear regressions together with the related tests. Recently, however, several factors (e.g., the size, complexity, and heterogeneity of financial data; the increasing complexity of regulatory requirements; the need to develop and deploy data products; and the importance of speed and efficiency) have increased the need for scalable technologies, which in turn calls for additional skills and requirements in the data analysis project (Varian 2014b; Dempster et al. 2018).

With novel and complex datasets, scalable technologies, and robust analytical methods, FDS promises to provide a wider range of opportunities and problem-solving methods for practitioners and researchers in the financial field. Crucially, this poses a significant challenge: The increased number of technological, methodological, and data choices might lead to technological chaos and complexity when managing the various steps involved in an FDS project. To overcome this problem, the author of this article proposes a holistic approach to organize an FDS project. The article is divided in two main sections: First, the ecosystem of financial data is examined to illustrate the sources, types, structure, and features of financial data; second, a nine-part approach to FDS management is offered and explained in detail.

THE ECOSYSTEM OF FINANCIAL DATA

Data are at the very heart of financial markets both as an essential input to financial and business processes and as output from them. Because a wide range of financial activities leaves a digital trace, the financial data ecosystem has emerged to include various sources, types, and structures of data. This section explores in detail the structure of the financial data ecosystem.

Sources of Financial Data

Understanding how the data were collected is of crucial importance in the scientific analysis of data. There exist five primary sources of financial data, which are mostly observational data.[4] First, several countries impose mandatory reporting requirements on firms regulated by government agencies. Mandatory filings are introduced to increase market efficiency and liquidity and reduce the cost of capital for firms (Admati and Pfleiderer 2000; Goldstein and Yang 2017). In the United States, for example, all companies, domestic and foreign, are required to regularly disclose information about their performance and operations to the Securities and Exchange Commission (SEC). Various types of filings are expected from each company. For example, the K-10 report provides comprehensive information about the performance of the company and is submitted within 90 days of the end of each fiscal year.[5] Data reported to the SEC are publicly available and can be accessed and downloaded for free via the EDGAR database, which maintains a vibrant collection of data forms filed by companies.

The main advantage of publicly collected data is that they are available regularly in a variety of frequencies and forms and are free to access. The main challenge, however, is the unstructured format of the data, which requires a lot of preprocessing and cleaning before the data can be analyzed. On the other hand, from a regulatory perspective, it might be nontrivial to determine the optimal level of disclosure precision to impose on firms (Admati and Pfleiderer 2000). Asking firms to disclose too much information might benefit market efficiency and liquidity, but it also might harm competition because firms would have more information about their competitors (Admati and Pfleiderer 2000; Goldstein and Yang 2017).

[4] In observational data, the units/subjects observed are not explicitly involved in the recording of the data. Experimental data in finance are rare owing to the ethical and monetary concerns that might characterize financial experiments.

[5] For a complete list of the SEC filings, check https://www.sec.gov/forms.

In addition to national bodies, financial data are also collected by international bodies, whose members are from several countries. For example, the Bank of International Settlements collects a wide range of financial data on international financial markets such as derivatives, debt markets, credit markets, liquidity indicators, and exchange rates.

The second source of financial data is exchange and trading facilities. A significant number of standardized financial products, such as stocks, derivatives, and bonds, are traded on exchange platforms such as NYSE and NASDAQ. Data about the trading and quotation of each product are collected and managed by the exchange. It should be noted, however, that a large number of financial products, especially derivatives and bonds, are traded over the counter (OTC), in which trading happens off exchange via a broker–dealer network. Because OTC trading is not conducted centrally, data on OTC transactions are often unavailable, inaccessible, or hard to collect.

Third, a valuable source of financial data is databases produced and curated by commercial data providers. Companies such as Thomson Reuters, Bloomberg, Standard & Poor's, and Capital IQ provide subscription-based access to a rich set of financial data that are frequently used in academic research and market analysis (see Exhibit 1 for a list of the most frequently used financial vendor datasets). Data providers collect data from regulatory agencies, news wires, company websites, and company reports and through voluntary reporting agreements with firms.

The main advantages of commercial databases are (1) they are often provided in a structured format, enriched with vendor-specific identifiers that better reflect the structure and dynamics of the markets; (2) they are easy to access and download; and (3) they come with documentation providing additional metadata. Nevertheless, several disadvantages might exist. For example, financial statement data retrieved from vendor databases are likely to be in a standardized format. Thus, if a company uses a unique accounting method to record its business operations, such information is likely to be lost in the data; this, in turn, might affect the analysis. The same is true when data are collected at the international level, where accounting standards might differ. Another downside of vendor data is the potential for biases, which often arise when the database is constructed by the voluntary reporting of entities that are not required by law to disclose their operations to the public.[6]

Data vendors provide data with different frequencies, usually through a financial data feed, which delivers data via an API or as raw files. Three types of data feeds exist: real-time feeds, in which data are pushed to the client as soon as they are available, using a software development kit or technologies like WebSockets; intraday data feeds, which deliver data during the day, usually every 15 minutes; and historical data feeds, which provide access to a long snapshot of past financial data.

The fourth source of financial data is alternative data, which can originate from sources such as social media, internal business processes, and websites. Factors such as the widespread use of social media, the adoption of mobile and Internet banking, and increasing regulatory requirements have translated into more data being generated and stored both within and outside the traditional context of financial institutions.

Finally, financial data can be artificially generated by financial scientists and institutions via simulations. This is a common practice in risk management, in which scenario analysis and market sensitivity simulations are performed to understand the reaction of the system or a variable to specific market changes.

Structure of Financial Data

Financial datasets come in a variety of structures and types, and they possess unique features that distinguish them from other types of data. In this section, the author will briefly examine the variety of data structures that can be encountered in an FDS project.

Financial time-series data. The most common structure of financial data is a *time series*. A time series is a collection of historical records of one or more variables, observed with specific frequency through time. Time series are ordered by time and can be expressed with the notation

$$X = \{X_t, t \in T\}$$

[6] For example, hedge fund datasets such as the Lipper Hedge Fund Database (TASS) are constructed by adding hedge funds that voluntarily decide to become part of the database. However, a hedge fund might decide to report in one period and skip the next one, perhaps because performance was bad and it fears investors will leave. Such behavior might introduce biases in the datasets that can distort the analysis.

EXHIBIT 1
A List of the Most Well-Known Commercial Datasets in Finance

1. CRSP
Provider: Center for Research in Stock Prices (University of Chicago)
Description: Accurate, survivor bias–free historical stock market data

2. COMPUSTAT
Provider: Standard & Poor's
Description: Fundamental and financial data on active and inactive firms worldwide

3. IvyDB
Provider: OptionMetrics
Description: Accurate end-of-day option prices as well as calculated implied volatilities and greeks

4. Loan Pricing Corporation DealScan
Provider: Thomson Reuters
Description: Rich and reliable information on the global syndicated lending market

5. Trade & Quote (TAQ)
Provider: New York Stock Exchange
Description: High-frequency data on trades and quotes for all issues traded on NYSE, NASDAQ, and the regional exchanges for a single trading day

6. Mergent Fixed Income
Provider: FTSE Russel
Description: Data on a large number of corporate, corporate medium-term note, supranational, US agency, and US Treasury debt securities

7. Trading Advisor Selection System (TASS)
Provider: Tremont Capital
Description: Comprehensive information on the hedge fund industry

8. Ancerno (Abel/Noser)
Provider: Abel Noser Corp.
Description: Institutional trading data

9. Institutional Brokers' Estimate System (I/B/E/S)
Provider: Thomson Reuters
Description: Stock analyst estimates on future earnings of US publicly traded companies

10. RavenPack News Database
Provider: RavenPack
Description: A primary source of news data

where T is an index set. Examples of financial time series are prices, returns, quotes, interest rates, volatility, and dividends. Such quantities can be quoted directly in the markets or inferred from other quantities.

Financial time series possess a number of temporal empirical characteristics that distinguish them from other types of economic and other data types (Cont 2001; Sewell 2011). Additionally, because of market efficiency and arbitrage forces, observed trends and regularities in financial time series can emerge, change, and eventually disappear in time (Malkiel 2003). This, in turn, translates to a higher noise component in financial time series (Tsay 2010).

Financial cross-section data. A cross-section data structure contains observations of one or more variables for multiple entities/subjects observed at one point in time. Very often, the selected subjects tend to share common features.[7] Formally, a cross section has the following form:

$$X_{i,t=\tau},\ i = 1, \ldots, N$$

[7] For example, to analyze R&D performance across firms, the cross section might include only those firms that possess an R&D department.

where i is the subject index and $t = \tau$ indicates a fixed point in time. Unlike time series, the ordering of cross-section data is irrelevant.

Typically, the main goal of using cross-section data is to perform point-in-time analysis to understand the differences between the selected subjects. In finance, this often involves the analysis of differences in performance across stocks, as well as heterogeneity in consumer and firm behavior. It should be noted, however, that cross-section data are not adequate if one wants to analyze the intertemporal effect of some variables (e.g., interest rate) on individual differences.

Financial panel data. A data structure that has both time-series and cross-section components is called a *panel* dataset. Formally, a panel has the following form:

$$X_{it}, i = 1, ..., N; t = 1, \ldots, T$$

where t is the time dimension, and i is the unit dimension. Panel data are used in many practical applications in finance, such as portfolio choice, factor analysis, asset pricing, and corporate finance (Fama and French 1992; Zhang and Zhou 2004; Koop 2005; Petersen 2009). Compared to cross-section and time-series data, panel data are more informative and present richer content that allows for (1) examining a wide range of complex and dynamic phenomena, (2) distinguishing between fixed and random effects, and (3) reducing the risk of collinearity (Baltagi and Song 2006).

Financial panel datasets can vary in the cross-section (N) and time-series (T) dimensions. In asset pricing, panels with long time series are often encountered, whereas in corporate finance, panels with a large cross section (N) are more common. A characteristic feature of panel data is their unbalanced content. A balanced panel is a dataset in which an observation for each unit is available at each period in the panel. Financial panel data can involve a variety of changes and dynamics over time that the researcher must take into account; for example, identifiers might be changed or reassigned, companies might change their business model, new companies might be born (initial public offerings or spinoffs), and others might disappear (bankruptcy, merger, or takeover).

Complex financial data structures. A characteristic feature of the previously discussed data structures (time series, cross section, and panel) is that they can be stored in a column-oriented format, in which data are arranged in rows and columns. Data can be stored in a variety of ways inside tables, and the tables can have relationships with each other. The way data and tables are organized in a database is called the *data model*. The combination of the data model with the data itself and any metadata is known as the *database schema*. The most common data model is the *relational* model, which uses a schema that stores data in different tables (called relations), with a key used to identify each row. To query and manipulate a relational database, structured query language (SQL) is often used. In many cases, people refer to a relational database as an SQL database.

Recently, new types of data have been employed in finance whose structure cannot be represented merely as column-oriented time-series, cross-section, or panel data. Examples include document-oriented data, graph data, and key-value data. To this end, new data models have emerged to store and structure these new data types. These data models are often known as *NoSql* databases, which means "not only SQL" (Varian 2014b). Contrary to SQL databases, NoSql databases are schemaless, meaning they do not impose rigid requirements on how the data should look or be arranged.

Types of Financial Data

The different structures of financial data discussed in the previous section can represent a variety of underlying phenomena. This section elaborates on the most relevant phenomena that can be represented by financial data.

Fundamental data. Fundamental data, sometimes called financial statements data (White, Sondhi, and Fried 1994) or corporate finance data (Flannery and Hankins 2013), are formal records that convey information about the performance (e.g., profits, costs, and sales), operations, and asset/liabilities structure of a firm.

The firms themselves often release fundamental data, specifically to regulatory agencies. The frequency of release of fundamental data is low, ranging from quarterly to yearly. To guarantee high-quality data, companies hire accounting firms to audit the data.

An essential feature of fundamental data is that they are often published with a *lapse*, meaning data for one time period are published at a later point in time (López de Prado 2018). Equally important features of fundamental data that might cause data consistency issues are data *backfilling* and *reinstatements* (Fabozzi 2012;

López de Prado 2018). Data backfilling happens when a firm's data are added to a database, and its history, which was not available previously, is also added. Reinstatements of data happen when a firm's initial data are revised and corrected, and the new corrected data overwrite the old version.

Transaction data. Transaction data, also called market data, encompass information about trading activities and security transactions that take place in a stock exchange (e.g., NYSE or TYO) or trading venues (e.g., BrokerTec Europe, Ltd. and MarketAxess).

Stock prices are the primary example of transaction data and they come in four types: transaction prices (the actual price of a transaction), firm prices (already arranged prices), indicative prices (for informational purposes only), and settlement prices (for daily margin calculations) (Holton 2003). Transaction prices are the most used in financial analysis and are recorded as open, close, minimum, and maximum prices.

Crucially, in theory, transaction prices are supposed to be recorded at a regular time interval, such as daily, weekly, or monthly. In practice, however, this is not the case: Transaction prices tend to be *nonsynchronous*, meaning different financial instruments trade at different frequencies and times during the day (Campbell, Lo, and MacKinlay 1997; Holton 2003; Tsay 2010). The nonsynchronous nature of financial data can raise issues during analysis, such as a lack of independently and identically distributed observations, absence of clear seasonality, and difficulties in forecasting random transaction times (Campbell, Lo, and MacKinlay 1997).

Another important feature of transaction data is *price discreteness*. Prices of listed stocks do not change continuously; instead, there is normally a minimum tick size allowed for price change (Engle and Russell 2004). Price discreteness can be beneficial for the functioning of markets (Anshuman and Kalay 1998), but it might cause problems when measuring price characteristics (e.g., volatility or dependence) that might require finer granularity relative to the tick size (Engle and Russell 2004).

Finally, a distinguishing feature of transaction data is their *high frequency*. To give an overview of high-frequency data, consider the TAQ database. TAQ provides data on all trades and quotes for all issues traded on NYSE, NASDAQ, and the regional exchanges for a single trading day. Historical data are available from 1993 to present. Between January 1993 and September 2003, the Daily TAQ data were available in non-millisecond timestamp (HHMMMSS). From October 2003 to July 24, 2015, TAQ data were offered with a millisecond timestamp. The Daily TAQ data are present with a microsecond timestamp from July 27, 2015 onward for (UTP) NASDAQ; all other exchanges in the Daily TAQ are presented in microseconds from August 3, 2015 onward. The Daily TAQ data are present with a nanosecond timestamp as of October 24, 2016 for (UTP) NASDAQ. All other exchanges in the Daily TAQ have a microsecond timestamp appended with three trailing zeros as of October 24, 2016 (IEC 2017).

Alternative data. An increasingly relevant type of data, employed extensively in the investment process and market analysis, is alternative data (Karppi and Crawford 2016; Guida 2018; López de Prado 2018). Alternative data can be produced outside the traditional context of financial institutions (e.g., news, social media feeds, web search data, and satellite images) or generated internally by the firms (e.g., client transactions, purchase behavior, and corporate operations).

Contrary to the more conventional datasets in finance, alternative datasets are not overly exploited and can offer unique and promising opportunities for those who collect and clean them. Nevertheless, it should be noted that the scope of alternative data might be more limited to specific stocks/events, in contrast to the universal application of conventional datasets (e.g., when constructing factors). In addition, challenges such as costs, privacy, and data cleaning workload can be encountered with alternative data. Another distinguishing feature is that alternative data, such as news, can often be subjective, which requires particular attention to the context of the data when interpreting it (see Loughran and McDonald 2011).

News and social media are perhaps the most prominent type of alternative data. The widespread use of news in financial analysis is recent. In 1961, Mindell (1961) claimed that news is too ambiguous of a data type to be used in financial analysis. However, the significant increase in the availability of news data and efficient techniques for processing them (in particular, natural language processing) has allowed for reliable analysis of the effect of the news on financial markets (Heston and Sinha 2017). News data can arrive with a high frequency and can be obtained in real time. Moreover, news can be prescheduled and synchronous (e.g., regular firm releases) or unexpected and asynchronous (e.g., corporate

decisions) (Mitra and Mitra 2011). Examples of useful information that can be extracted from news include the tone of the news (positive, negative, or neutral), its strength (by deriving an impact score), its event class (what kind of news), and its novelty (see Mitra and Mitra 2011).

Financial network data. In the wake of the 2008 financial crisis, particular interest about the importance of interconnectedness in financial markets has emerged, mainly for risk management and financial stability purposes. Financial markets are an outstanding example of a networked system in which nodes can represent firms, financial instruments, or individuals. Among these nodes, links can be observed directly (e.g., cross holdings, Haldane and May 2011; interbank loans, Boss et al. 2004; and interlocking directorate, Battiston et al. 2010) or inferred from other quantities (e.g., similarity in portfolios based on their compositions or similarity of stocks based on pairwise correlations; Mantegna 1999; Battiston et al. 2010).

Studying financial networks can provide substantial input for understanding and predicting systemic risk, financial contagion mechanisms, risk concentrations, and too-central-to-fail financial institutions (Battiston et al. 2012; Glasserman and Young 2016; Das, Kim, and Ostrov 2019). Other types of financial data, such as fundamental data, might lack the power to uncover such phenomena or complex market structure (Mertzanis 2018). Moreover, traditional economic tools can fail to model and predict systemic risk; this can be overcome by using financial network data and tools from network science and complex systems (Battiston et al. 2016).

An essential and frequent issue when dealing with financial data is availability. Data on important network phenomena such as the interbank market, derivatives market, and exposures are mostly not disclosed to the public for privacy and confidentiality reasons. Instead, banks might only disclose aggregate information (e.g., Bank A has X amount of assets with the rest of the system). This might prevent the researcher from conducting reliable microanalysis. In such cases, various methods have been proposed to reconstruct the network from aggregate data (Cimini et al. 2015). On the other hand, even if detailed data were available, they often lack granularity. Low granularity of financial network data means that the data are partial (some of the actual links are present in the data, but the rest are missing). In this case, several methods can be employed to infer the missing links (Anand et al. 2018).

So far, this article has focused mostly on the financial data ecosystem and its features. To efficiently navigate this ecosystem and exploit the opportunities it offers, an FDS strategy is needed to organize the various steps and components involved. The rest of the article will offer a detailed account of how to implement such a strategy.

THE FDS STRATEGY

A distinguishing feature of modern data science projects is the multiplicity of parts and steps involved and the interdependencies between them. Without a strategy and plan, an FDS project might be subject to high uncertainty, nuisances, and inefficiency. In what follows, the author presents a holistic approach to efficiently organize an FDS project into nine interrelated steps: data culture, problem identification, technological stack design, data gathering, exploratory data analysis, data integrity, feature analysis, model development, and interpretation.

An FDS strategy is both iterative and collaborative. It is iterative in that it involves several back-and-forth episodes of testing new technologies, ideas, methods, and features to improve the overall performance of the project. Collaborative here means that, throughout the life cycle of a project, several individuals with various backgrounds and expertise are involved. The next sections will elaborate on each of the nine parts in detail.

Data Science Culture

The idea that organizations/decision-makers should rely on information conveyed by data has been increasingly gaining popularity. Such practice is often called *data culture* or *data-driven decision making*. In this article, the author claims that a data culture is needed to establish the standards and rules required to guarantee sound ethical and methodological practices throughout the lifecycle of an FDS project.

On the methodological side, a relevant practice in financial research concerns the weights to assign to both the theory/experience-driven and data-driven aspects of the project. A pure data-driven culture toward scientific analysis, with no ex ante economic foundations, is likely to be a risky and unreliable approach (Arnott, Harvey,

and Markowitz 2019).[8] Equally risky is finding ex post economic foundations (after the data analysis). With the wide variety of algorithms and data manipulation techniques available today, it is almost guaranteed that a pattern will be found in the data. Without ex ante foundations/hypotheses, nonexperimental fields like finance can easily fall into the trap of *data snooping* (also called *data dredging*), in which the analyst finds what seems significant but is, in fact, a spurious pattern in the data (Lo and MacKinlay 1990; Bailey et al. 2017; López de Prado 2018). Additionally, failure to understand or implement the right analytical methods might lead to costly errors when introducing a data product to the market (e.g., fire sales, flash crashes, botched initial public offerings, and catastrophic algorithmic trading errors; Lo 2017).

On the ethical side, a data-oriented culture must always take into consideration customer protection measures such as privacy and secrecy and be aware of potential discrimination and biases both in the data and the products created from them. Neglecting these aspects in FDS projects might lead to biased and discriminatory behavior being automated in products such as robo-advisors (Baker and Dellaert 2017) and credit scoring (Hurley and Adebayo 2016). Last but not least, a sound ethical practice that is worth encouraging, especially in financial research, concerns the disclosure of detailed/summary information on both the successful and failed trials that a scientist performed in the data analysis process (Harvey 2017; Arnott, Harvey, and Markowitz 2019; López de Prado 2019).

Problem Identification

The foremost driver of an FDS project is the identification of a problem or research question that requires scientific analysis of financial data. In academia, researchers might be interested in a variety of financial phenomena such as market efficiency, manager performance, factor analysis, and various corporate finance topics. In industry, financial firms and professionals might need to construct investment portfolios, automate trading, detect financial fraud, predict customer behavior, and recommend products, to name a few. A common element in all of these problems is the need for data.

The practice of identifying what constitutes a significant/interesting problem and what data are needed must be nurtured through the data science culture. Theoretically speaking, there is no recipe for what constitutes an interesting question. However, because problems like financial ones are human-made, it is natural to assume that humans relying on intuition are the best judge of what constitutes an interesting hypothesis/question (Blei and Smyth 2017; Arnott, Harvey, and Markowitz 2019).

Technological Stack Design

To materialize an FDS project, an essential requirement is the technologies needed to interact with the data and ensure project sustainability. In this article, the term *data science technological stack* is used to refer to the combination of computation and storage infrastructure, programming languages, frameworks, libraries, and paradigms used in the development of an FDS project. The technological stack plays a primary role in driving the efficiency and effectiveness of every step in an FDS project. Poor management of the technological stack might lead to technological chaos and nuisances that hold back productivity.

Practically speaking, there is no single optimal recipe for choosing the technological stack. Several trade-offs are involved that require careful thinking. To mention a few, consider the choice of storage and computation infrastructure. From one perspective, two main options are available: *dedicated servers* and *cloud solutions*.[9] Both options have their pros and cons. Cloud solutions offer on-demand scalability (e.g., adding more CPU, GPU, RAM, or storage), resilience (recovering after a failure), and absence of maintenance or configuration on the client's side (Hashem et al. 2015). The downsides

[8] Contrary to common perception, modern data science applications are not purely data driven. Take, for example, *robo-advisors*, which are automated online investment advisors. Robo-advisors are often mentioned as an example of a data-driven product, which is, in part, true; however, the principles upon which these solutions are built do also rely on theoretical findings in the financial literature, such as financial microeconomics, behavioral finance, and portfolio theory (https://www.betterment.com/resources/nobel-prize-winning-investing/).

[9] Simply put, cloud solution refers to the delivery of computing and storage services over the Internet, whereas dedicated servers refers to the practice of buying hardware and software and building an on-site storage and computation infrastructure. It should be noted, however, that an institution can buy physical infrastructure and create a private cloud. For more details, see https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/.

of the cloud are the high costs and the lack of control on the client's side. With dedicated servers, the user has full control over the data and the server configuration, but scalability and resilience are often harder to achieve.

For financial applications, a key criterion in technological stack design is performance. To achieve high performance, an essential paradigm that worth mentioning is *distributed systems*. A distributed system is a collection of computers, often called nodes, that communicate via a network by passing messages to each other; a distributed system appears to the human user to be a single, coherent system (Tanenbaum and Van Steen 2007). An important type of distributed system that is of particular relevance to finance is *high-performance computing* (HPC) (Dempster et al. 2018; López de Prado 2018). HPC is the practice of aggregating substantial computing power to process data and perform computations in parallel at a very high speed. The need to process real-time data and the computationally intensive nature of various financial operations are the main reason for use of HPC in finance (Dempster et al. 2018).

Another trade-off involved in technological stack design occurs when choosing the programming ecosystem. A programming ecosystem includes a core programming language and all the libraries, frameworks, and packages developed around it. Examples of programming ecosystems include open-source solutions such as Python and R and commercial products such as Matlab, all very common in finance.[10] Open-source solutions allow for flexibility in development and experimentation when compared with commercial solutions, but this comes at the expense of more vulnerability to unseen bugs that can affect computation and security.

Recently, there has been an evolution in a cloud-based technology that the author calls *unified analytics platforms* (UAP). A UAP leverages the power of the cloud and virtualization technologies[11] to offer a unified framework to work with data, files, and applications. Thanks to virtualization, UPAs allow for reusability of code and reproducibility of results, which are highly relevant issues in any data science project (Lyuu 2001; Carmichael and Marron 2017). Examples of UAP include business-oriented platforms, such as Amazon SageMaker, Google Cloud AutoML, and DataBricks analytics platform, and academic-oriented platforms such as Alphacruncher's Nuvolos platform.

Data Gathering

A distinguishing feature of modern data science is the greater emphasis on data gathering (Donoho 2017). In an FDS project, the gathering and preparation of the data are as important as the identification of the problem itself. Datasets can vary in their coverage and quality; therefore, identifying the right dataset can have a substantial impact on the analysis outcome (Chakrabarty and Trzcinka 2006).

The choice of dataset is primarily dependent on the nature of the problem. For example, to study yearly seasonality in stock markets (e.g., the January effect), low-frequency data would be sufficient. However, if the scientist is interested in detecting intraday patterns, stock data at a higher frequency are needed (Giot 2005). Some problems might require more than one type of dataset; for example, analyzing the effect of the news on financial markets would require a dataset for news and another for the stock market.

Depending on the subject trying to gather the data, factors such as privacy, confidentiality, and competition might constrain access to specific sources of data. For example, data on derivatives, risk exposures, and interbank lending are considered highly confidential, and market participants are likely to share it only with trusted parties such as regulators and central banks. Therefore, a crucial step in an FDS project is to ensure that access to the required data is feasible.

When the data needed for the analysis are identified and access to them has been secured, the next step involves the importing, formatting, and storing of the data for preprocessing. Data engineers often do this through a process known as *extract, transform, load* (ETL). An ETL process is a pipeline composed of multiple steps that involve (1) the extraction of data from one of multiple sources, using tools such as API, SKD, web scraping, or in-house tools; (2) the transformation of the data into a format that better fits the needs of the project (using a set of rules and functions); and (3) loading the data into a target storage place, which is often a database

[10] A project might involve the use of components from different ecosystems, which can communicate via APIs.

[11] Virtualization is an abstraction layer that allows multiple virtual environments or software to be hosted on the same physical system. With virtualization, it is possible to create a uniquely customized environment configured to execute a specific task and obtain the same outcome, regardless of the machine in which they are hosted.

(sometimes called data warehouse). Financial data ETL processes can be performed in stream (real time) or batch (historical) modes. An ETL process can involve several challenges, such as process complexity, heterogeneous data sources, scalability to large datasets, download limits imposed by data vendors, and errors.

Exploratory Data Analysis

Once the data have been prepared for analysis, the next step would be to carry out an initial investigation of the data through a process known as *exploratory data analysis* (EDA). EDA relies on graphical and statistical tools to understand the structure and features of the data, explore summary statistics, uncover general patterns, spot anomalies, visualize correlations, and test basic assumptions and hypotheses. EDA is of crucial importance to an FDS project because it can assist in several steps of the project, such as identifying the right problem and the right dataset(s), confirming data integrity, conducting feature analysis, and developing the model.

A variety of EDA tests can be performed on financial data. For example, an EDA of financial time series might involve the decomposition of the time series into temporal components such as trend or seasonality, level component (average value), and noise (random component). Such a decomposition would be useful in investigating empirical features of financial data such as the low signal-to-noise ratio[12] and seasonal patterns such as the January effect, day-of-the-week effect, and turn-of-the-month effect (Malkiel 2003).

Equally important would be the investigation of the so-called *stylized facts* of financial markets. For instance, stock data are thought to possess empirical properties such as lack of autocorrelation, non-normality of returns, volatility clustering, and scaling (Cont 2001). Another example of a stylized fact concerns financial network data, which often show a *core–periphery structure*, in which the network core is densely connected and the periphery is made of sparsely connected nodes that mostly link to the core.

EDA of financial data can be complicated by several factors. The main challenge is the presence of large and multidimensional datasets. With small datasets and a few features, the scientist can perform exploratory data analysis and quickly obtain useful information. With large and multidimensional datasets, both the time requirements and the complexity of EDA increase considerably (Einav and Levin 2014).

Data Integrity

As with all empirical fields, the quality of the final output of an FDS project depends significantly on the quality of the data employed in the analysis. This is a direct result of the *garbage in, garbage out* principle, which states that the quality of the output is determined by the quality of the input. To this end, in an FDS project, a variety of checks and tests should be performed to identify and eliminate potential errors, biases, and artifacts and to ensure data integrity.

Data errors. Data errors are digital entries that have been recorded erroneously and thus reflect an incorrect or meaningless value. The presence of data errors, even at low probability, can seriously damage scientific analysis and potentially lead to wrong decisions and conclusions.

Errors in financial data can emerge for a variety of reasons. According to Falkenberry (2002), voluminous data and asynchronous trading processes represent the primary source of data errors. Additionally, very intense trading is likely to cause errors when recording information. Finally, human errors might also be a source of erroneous data—for example, if the data are manually keyed (Holton 2003).

Examples of financial data errors include wrong decimal places, negative prices, extra or removed digits, wrong maturity dates of securities, inverted exchange rates, rounding, exchange of values, and many more (Rosenberg and Houglet 1974; Holton 2003; Ince and Porter 2006). Such errors might be obvious or subtle. Obvious data errors, such as a price jump from $1 to $1,000,000, can be easily recognized by computer algorithms.[13] More subtle errors, such as a price of $40, followed by two prices of $4, and then a fourth price of

[12] This behavior is mainly caused by arbitrage effects and the speed of news generation and consumption in financial markets (Meinl and Sun 2015; López de Prado 2015; Easley, López de Prado, and O'Hara 2016; Rasekhschaffe and Jones 2019).

[13] A variety of computer algorithms can be employed to recognize data errors. These algorithms may range from pattern recognition techniques (e.g., distance-based techniques) to statistical inference and arbitrage relationships (e.g., interest-rate parity or put–call parity) (Holton 2003).

$40 might require more in-depth investigation. A frequent and powerful approach to correct financial data errors is *cross-dataset validation*, which consists of comparing data from one source against an alternative data source that records similar data (Wood, McInish, and Ord 1985; Ince and Porter 2006).

An essential challenge in financial data error filtering, in particular with high-frequency data, concerns the trade-off between *underscrubbing* and *overscrubbing* (Falkenberry 2002; Brownlees and Gallo 2006). With too much filtering, there is the risk of losing connection with reality and altering the statistical properties of the data, whereas too little filtering is very likely to yield an unstable dataset. Data scientists should pay careful attention to how to achieve a balance in this trade-off.

Data biases. Differing from data errors, data biases are distortions in data that reduce the sample's representativeness of the population or phenomenon under examination. A large number of biases might occur in datasets (Mehrabi et al. 2019), but to give an idea specific to financial data, the author considers proprietary hedge fund data as an example (Fung and Hsieh 2000).

Hedge fund data are often collected via the voluntary reporting of hedge fund managers to commercial data vendors. This is often done to highlight performance and attract capital. The hedge fund industry is notably a risky business; therefore, the probability of failure is high. If only the failed or poor-performance funds are systematically removed from the database and the winners are kept, this might give rise to *survivorship bias*. It is a bias because, when studying the return of such funds, the researcher might fall into the trap of overstating the success of hedge fund management. On the hedge fund side, a manager might decide to stop reporting to the database either because it is performing poorly and wants to hide results or it is outperforming and does not want additional inflow of capital. This behavior would cause *self-selection bias*. A bias related to survivorship and self-selection is *backfilling bias*, also known as *instant history bias*. Backfilling bias happens when a fund is added to a database and its historical data are also added. This might cause bias because only funds that have had a good history tend to decide to join the database, thus skewing the past performance of the hedge fund industry. Finally, a crucial bias is called *look-ahead bias*. This bias happens when a study uses data that would not have been available during the time period being studied. For example, assuming that a fund would survive until the end of the sample period, which is data that are not available at the time of the study.

Controlling for biases in datasets is not trivial. However, the first step is to understand how the data were generated and recorded. In the literature, a variety of debiasing techniques have been proposed (Mehrabi et al. 2019). The use of *cross-dataset validation* could be a valuable tool, where the goal is to check whether similar patterns can be detected in datasets that contain similar or related data. On the commercial data vendor side, efforts have been made to detect and correct biases in the datasets. For example, TASS started recording data on defunct funds in a separate database, called Graveyard. Another example is CRSP, which had to account for delisting returns bias in its data.

Data availability. A common problem in financial datasets is data availability or data incompleteness (Kofman and Sharpe 2003; Fabozzi 2012). Incomplete data might arise for a variety of reasons. First, if disclosure of certain financial items is not mandatory by law, data for those items might be unavailable for some of or the whole sample period. Second, as illustrated previously in this article, factors such as nonsynchronous trading imply that data for an actively traded stock will be available more often than data for a nonactive stock. Third, firms might release data at different frequencies and with different details. Fourth, data might only be available at the aggregate level, or different variables might be combined. Fifth, certain types of data might be available only for a subset of the entire dataset. For example, data on patents might be available for many sectors, but for sectors like finance in which patenting is not a common practice, data are often lacking.

A variety of methods exist for dealing with incomplete financial datasets. Each method has its pros and cons. According to Kofman and Sharpe (2003), *likewise deletion*, in which observations with incomplete data are removed, is the predominant approach. The *omitted variable* method instead removes variables that have incomplete data. The third approach is *imputation*, in which missing data are estimated and then treated as observed.[14]

Alternatively, a powerful technique for dealing with incomplete data is *data matching*, also known as *entity*

[14] In the case of nonsynchronous and irregular data, the method of *interpolation*, which can be treated as an imputation technique, is often used to convert irregularly spaced data to regularly spaced data (Brownlees and Gallo 2006).

resolution or *record linkage*. According to Christen (2012), data matching is the task of "identifying, matching, and merging records that correspond to the same entities from several databases."[15] If one data source contains observations or variables that are missing in another source, then matching the two datasets should result in a richer dataset. The practice of data matching in finance is widespread; for example, CRSP stock data are frequently merged with COMPUSTAT to obtain fundamental data.

Matching of financial data is often done in two ways: First, if the same identifier is present in two datasets, this can be used to match similar records (e.g., using CUSIP or Ticker). The main challenge with this approach is that these identifiers are not permanent, and in the case of tickers, they can be reassigned. Moreover, specific datasets retain *header* data only, where historical records are not retained.[16] The other approach is to create a *link table*, which is a table that matches different identifier(s) in one dataset with identifier(s) in another dataset. A well-known example of the link table is the link history table contained in the CRSP–COMPUSTAT merged database.[17]

Feature Analysis

Once data integrity checks have been performed, the next step is to carry out *feature analysis* to identify and prepare the features to be used in the model.[18] The performance of an FDS project is crucially dependent on having the right feature space that correlates with the target variable. Feature analysis includes three (iterative) processes: feature *engineering*, *projection,* and *selection*.

The practice of selecting the features to be included in the model is known as feature selection. Recently, the importance of feature selection has increased as the dimensionality of data has grown (Kuhn and Johnson 2013). High dimensionality could be useful because it enriches the dataset and might help in obtaining unbiased estimators (e.g., by avoiding omitted variable bias). However, high dimensionality could also mean a higher risk of collinearity, increased model complexity, and less intuitive understanding of the data and analysis outcome. In addition, research has shown that for a learning algorithm to produce robust predictions, the number of observations should grow exponentially with feature dimensionality. This phenomenon is often known as the *curse of dimensionality* (Verleysen and Francois 2005).

High dimensionality can often be encountered in financial data. For example, in portfolio analysis, the number of parameters (which we can treat as features) can easily exceed the sample size.[19] Moreover, because of the limited scope of financial data,[20] escaping the curse of dimensionality can be quite challenging. Other examples of high-dimensional financial data are the large number and variety of investment funds (e.g., exchange-traded funds), asset managers, indexes, financial ratios, investment and trading strategies, product purchase combinations, and various security combinations.

Notably, although the number of features in financial analysis can be high, research has shown that there could be universality in the set of features that affect financial markets (Fama and French 1992; Sirignano and Cont 2019). Research on multifactor models has revealed that financial asset returns can be decomposed into returns derived from exposure to common factors[21] and an asset-specific component (Fama and French 1992). Such universality has played a vital role in reducing the number of parameter estimates in financial analysis (Mitra, Mitra, and diBartolomeo 2009). Finally, in addition to these finance-specific techniques, researchers in statistics and machine learning have worked on a variety of methods for feature selection.[22]

[15] The author noted that the practice of cross-dataset validation mentioned earlier in this article is performed via data matching.

[16] For example, if a company had its headquarters in one city for 10 years and then moved it to another city, some datasets will show only the second address throughout the entire history of the company.

[17] The CRSP–COMPUSTAT merged database is a linking database created to facilitate the merging of stock data from CRSP with fundamental data from COMPUSTAT.

[18] A feature is a variable that represents raw data.

[19] For example, the number of parameter estimates required for mean–variance optimization based on the S&P 500 universe is 125,750: 500 expected return estimates, 500 volatility parameter estimates, and $[500 * (500 - 1)]/2 = 124{,}750$ correlation parameters.

[20] Apart from high-frequency data, financial data are often limited in scope, and the sample size that can be obtained is constrained by a limited history (Arnott, Harvey, and Markowitz 2019).

[21] Factors are common characteristics of assets, representing different risk aspects, that can be employed to understand the cross-section differences in asset performance.

[22] These methods are often grouped under three classes: In *filter methods*, a feature is selected based on its score in a statistical test against the target variable (e.g., Pearson correlation and chi-square). *Wrapper methods* select features based on a search process that

Although the process of feature selection reduces the dimensionality of the raw data, in some cases, scientists might exploit dependencies among the features to reduce dimensionality by transforming data in the original high-dimensional space into a space of lower dimensions. This practice is often called feature extraction, feature projection, or simply dimensionality reduction. Examples of dimensionality reduction techniques are *principal component analysis* and *factor analysis*.[23] The main advantage of feature extraction is the higher computational performance, and the main disadvantage is the loss of interpretability of the data, given that the new projected data are not what was produced by the real system.

In some cases, the nature of existing features might be inadequate or unsatisfactory for representing the problem or phenomenon at hand. In such a case, data scientists use intuition and domain knowledge to create new features via the process of feature engineering (Klaas 2019; Rasekhschaffe and Jones 2019). Analysis of financial data often involves feature engineering (Koop 2005; López de Prado 2018). For example, in time-series analysis of stock markets, analysts make sure they use adjusted stock prices, which account for corporate operations such as dividend distributions, stock splits, spinoffs, and mergers and acquisitions. Other features engineering examples include time-series differencing, log transformation, and price to return transformation.

The practice of feature engineering should be treated with caution. For example, price differencing is often performed to guarantee stationarity or normality in the data, which is desirable when employing inferential statistics (Campbell, Lo, and MacKinlay 1997; Ruppert 2011). However, price differentiation would erase memory from the time series, thus reducing the signal in the data and increasing the risk of false discoveries. In such a case, methods have been proposed to achieve stationarity while preserving as much memory as possible (López de Prado 2018). Finally, given the large number of possible combinations, a balance in feature engineering is required. It is essential to make sure that data scientists do not fall into the trap of never-ending feature engineering.

Model Development

After thorough exploration and pre-preprocessing of the data, the next crucial step is to choose, train, and evaluate a model to solve the scientific problem at hand using a specific algorithm. A large number of factors exist that can affect the choice of a model or an algorithm. The goal of this section is to discuss a short, nonexhaustive, list of rules on how to perform model development.

First, before developing a model, a remarkable result to take into account is the *no free lunch theorem* (NFL) for optimization (Wolpert and Macready 1997). The NFL theorem states that there is no single algorithm that is universally the best: Better performance observed for one dataset/problem is likely to be compensated by worse performance on a different dataset or problem. Add to this the fact that financial markets are dynamic and, in many cases, efficient (at least locally in time) (Timmermann and Granger 2004). As a consequence, firms that want to try to exploit market inefficiencies in different time and space pockets need to continually develop different, specifically tuned predictive models because the old models will be spotted by arbitrage and exploited by imitators.

Second, the choice of the model depends primarily on the purpose of the analysis. In this regard, an important distinction holds between *generative* models, concerned with modeling and understanding the data-generating process, and *algorithmic* models, in which the main goal is out-of-sample prediction (Breiman 2001).[24] Generative models tend to give additional importance to properties such as unbiasedness, efficiency, and consistency. In practice, the complexity of the data might require the use of models that, without serious consequences, violate some of these assumptions. For example, models that intrinsically reduce feature dimensionality (e.g., LASSO) are automatically biased.

Third, model selection can be affected by the number and nature of ex ante model assumptions. In financial econometrics, it is typical to start the

iteratively adds or removes features (e.g., recursive feature elimination). Finally, *embedded methods* refer to those models that intrinsically conduct automatic feature selection (e.g., LASSO).

[23] Factor analysis here is a statistical method that is different from financial factor modeling, as by Fama and French (1992).

[24] This distinction is perhaps what differentiates most of statistics/econometrics (generative model oriented) from machine learning (prediction model oriented).

modeling with ex ante assumptions on the functional shape. In this regard, data scientists often develop a *parametric* model to understand the data. In a parametric model, such as univariate linear regression, a dependent variable y is modeled as a function of an independent variable x—that is, $y = f(x) + \varepsilon$ where the form of $f(x)$ is assumed to be known (i.e., a linear function of the form $y = \theta_0 + \theta_1 + \varepsilon$). Instead, in a nonparametric approach, the shape of $f(x)$, together with its parameters, is not known in advance. A nonparametric model, such as *Gaussian process regression*,[25] finds a distribution over all candidate $f(x)$ that are consistent with the given data. As more data are obtained and used, a nonparametric model can update its shape and parameters. Parametric models are simpler, less data demanding, and less computationally expensive than nonparametric ones, whereas nonparametric models promise better flexibility, scalability to large datasets, and performance. Both parametric and nonparametric models have been extensively used in finance.

Fourth, the distribution of the data is a decisive factor in developing the model. In financial research, it is typical to assume that data are normally distributed (Wood, McInish, and Ord 1985; Tu and Zhou 2004), even though it is a widely accepted fact that financial data exhibit Leptokurtic and, in many cases, asymmetric behavior (Kon 1984; Mills 1995). Distributional complexity often translates into model complexity. For example, less complex data tend to present a linear structure; in such cases, parametric methods such as logistic regression and linear discriminant analysis are adequate choices. Linear models are fast and straightforward to train, but they tend to underperform (by underfitting) when modeling nonlinear data and outliers (Campbell, Lo, and MacKinlay 1997; Flach 2012).

With nonlinear models, it is possible to model any functional form that fits the data. In this regard, the methods of *deep learning* have recently gained significant influence in all fields, including finance. The idea of deep learning is to learn a hierarchical representation of the data, which is best illustrated through the use of neural networks with multiple layers. Neural networks derive their power from the repeated aggregation of nonlinear functions at multiple levels. For this reason, and following the *universal approximation theorem*, neural networks are said to be able to approximate any function (Aggarwal 2018).

In addition to deep learning, a robust modeling approach is *ensemble learning*, in which multiple models (often called weak learners) are combined after being trained to solve the same problem. The idea that a combination of alternative forecasts would outperform a single forecast has always been an important subject of research (Bates and Granger 1969) and recently has gained increasing attention in financial analysis (Huang, Nakamori, and Wang 2005; López de Prado 2018; Rasekhschaffe and Jones 2019).

Once a model has been chosen, the next step is to train the model using the data sample. Training often happens by dividing the data into a *training*, *validation*, and *test* sets. The model is initially fit on the training set. Successively, the validation set provides an unbiased set to evaluate the performance (overfitting/underfitting) and allows for *hyperparameter*[26] tuning. Finally, the best-performing model chosen in the validation phase is used to evaluate the *generalization error* using the test set. In generative models, further analysis can be performed for parameter estimation and significance.

Model evaluation is done using a performance metric. Depending on the problem at hand, different levels of performance can be acceptable. In industry, model accuracy directly correlates with the business value of the product that uses the model. Unlike some other fields,[27] the cost of model errors in finance might be consequential, given the monetary nature of financial applications. For these reasons, the FDS project manager should set performance targets, estimate the business value of model accuracy, and evaluate alternatives in case model performance hits a plateau.

Interpretation and Communication of Results

In any FDS project, understanding and communicating the model and its outcome are of utmost importance. Typically, on a practical level, a trade-off exists between model interpretability and complexity.

[25] See De Spiegeleer et al. (2018) for an application of Gaussian process regression in finance.

[26] A hyperparameter is a parameter whose value is set before training the model (e.g., the regularization parameter in LASSO).

[27] For example, when you call an Uber, you get an estimate of the car's arrival time, which is the output of a predictive model. An increase in the accuracy of the estimate of one or a few seconds is likely not to have a large impact on the number of orders and trips.

Straightforward interpretability is not always achievable owing to the complexity of real-life data and problems that require more sophisticated models (Kuhn and Johnson 2013). Nevertheless, seeking interpretable models should be treated as one of the project goals for a variety of reasons.

First, interpretability assumes relevance whenever there is incompleteness in the problem formalization (Doshi-Velez and Kim 2017). For some problems, prediction might only be part of the answer we are seeking. To complete the answer, complementary factors such as human judgment, domain knowledge, fairness, safety, reliability, and robustness might be needed. FDS projects are prone to costly mistakes, biases, false discoveries, and fast market dynamics that can be better accounted for via better model interpretability.

Second, interpretability can constitute an advantage for market participants. Because market participants are accessing more of the same data, it is reasonable to believe that the advantage shifts to those who have higher means of communication and interpretation (Beunza and Stark 2004).

Finally, failure to understand or interpret data science models might lead to a lack of trust by decision-makers, who will potentially forgo the benefits of adopting certain useful models (Miller 2019). A suitable example of this issue is the capital asset pricing model (CAPM),[28] which has achieved high popularity among professionals and decision-makers. Among the reasons for the high popularity of CAPM is its intuitive interpretability: It simply states that to achieve a higher return, one has to assume more systemic risk.

CONCLUSION

FDS is a promising field, offering a wide range of opportunities for industry professionals and researchers. At the same time, the number and variety of components and steps involved in an FDS project have been continually increasing, thus generating the need to take an integrated view of FDS. To this end, the present article proposes a holistic approach to think, plan, and efficiently organize the various components involved throughout the life cycle of an FDS project. Adopting the proposed strategy would improve the productivity and reliability of an FDS project. The recommended approach is comprehensive, but it is not exhaustive. For some FDS projects, additional steps and components, tailored to the problem at hand, might be required.

[28] CAPM is expressed as $ER_i = R_f + \beta_i(ER_m - R_f)$, where ER_i is the expected return of an asset i, R_f is the risk-free rate, β_i is a measure of asset i systemic risk, and $ER_m - R_f$ is the market risk premium.

REFERENCES

Admati, A. R., and P. Pfleiderer. 2000. "Forcing Firms to Talk: Financial Disclosure Regulation and Externalities." *The Review of Financial Studies* 13 (3): 479–519.

Aggarwal, C. C. *Neural Networks and Deep Learning.* Cham: Springer International Publishing, 2018.

Anand, K., I. van Lelyveld, Á. Banai, S. Friedrich, R. Garratt, G. Hałaj, J. Fique, I. Hansen, S. Martínez Jaramillo, H. Lee, J. L. Molina-Borboa, S. Nobili, S. Rajan, D. Salakhova, T. C. Silva, L. Silvestri, and S. R. Stancatode Souza. 2018. "The Missing Links: A Global Study on Uncovering Financial Network Structures from Partial Data." *Journal of Financial Stability* 35: 107–119.

Anshuman, V. R., and A. Kalay. 1998. "Market Making with Discrete Prices." *The Review of Financial Studies* 11 (1): 81–109.

Arnott, R., C. R. Harvey, and H. Markowitz. 2019. "A Backtesting Protocol in the Era of Machine Learning." *The Journal of Financial Data Science* 1 (1): 64–74.

Bailey, D. H., J. Borwein, M. López de Prado, and Q. J. Zhu. 2017. "The Probability of Backtest Overfitting." *Journal of Computational Finance* 20 (4): 39–69.

Baker, T., and B. Dellaert. 2017. "Regulating Robo Advice across the Financial Services Industry." *Iowa Law Review*, 103713.

Baltagi, B. H., and S. H. Song. 2006. "Unbalanced Panel Data: A Survey." *Statistical Papers* 47 (4): 493–523.

Bates, J. M., and C. W. Granger. 1969. "The Combination of Forecasts." *Journal of the Operational Research Society* 20 (4): 451–468.

Battiston, S., J. D. Farmer, A. Flache, D. Garlaschelli, A. G. Haldane, H. Heesterbeek, C. Hommes, C. Jaeger, R. May, and M. Scheffer. 2016. "Complexity Theory and Financial Regulation." *Science* 351 (6275): 818–819.

Battiston, S., J. B. Glattfelder, D. Garlaschelli, F. Lillo, and G. Caldarelli. "The Structure of Financial Networks." In *Network Science*, pp. 131–163. New York: Springer, 2010.

Battiston, S., M. Puliga, R. Kaushik, P. Tasca, and G. Caldarelli. 2012. "Debtrank: Too Central to Fail? Financial Networks, the Fed and Systemic Risk." *Scientific Reports* 2: 541.

Beunza, D., and D. Stark. 2004. "Tools of the Trade: The Socio-Technology of Arbitrage in a Wall Street Trading Room." *Industrial and Corporate Change* 13 (2): 369–400.

Blei, D. M., and P. Smyth. 2017. "Science and Data Science." *Proceedings of the National Academy of Sciences* 114 (33): 8689–8692.

Bose, I., and R. K. Mahapatra. 2001. "Business Data Mining—A Machine Learning Perspective." *Information & Management* 39 (3): 211–225.

Boss, M., H. Elsinger, M. Summer, and S. Thurner. 2004. "Network Topology of the Interbank Market." *Quantitative Finance* 4 (6): 677–684.

Breiman, L. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.

Brooks, C., A. G. Hoepner, D. McMillan, A. Vivian, and C. Wese Simen. 2019. "Financial Data Science: The Birth of a New Financial Research Paradigm Complementing Econometrics?" *European Journal of Finance* 25 (17): 1627–1636.

Brownlees, C. T., and G. M. Gallo. 2006. "Financial Econometric Analysis at Ultra-High Frequency: Data Handling Concerns." *Computational Statistics & Data Analysis* 51 (4): 2232–2245.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay. *The Econometrics of Financial Markets*, vol. 2. Princeton, NJ: Princeton University Press, 1997.

Carmichael, I., and J. Marron. 2017. "Data Science vs. Statistics: Two Cultures?" *Japanese Journal of Statistics and Data Science* 1: 117–138.

Chakrabarty, B., and C. Trzcinka. 2006. "Momentum: Does the Database Make a Difference?" *Journal of Financial Research* 29 (4): 441–462.

Chen, W. S., and Y. K. Du. 2009. "Using Neural Networks and Data Mining Techniques for the Financial Distress Prediction Model." *Expert Systems with Applications* 36 (2): 4075–4086.

Christen, P. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. New York: Springer Science & Business Media, 2012.

Cimini, G., T. Squartini, A. Gabrielli, and D. Garlaschelli. 2015. "Estimating Topological Properties of Weighted Networks from Limited Information." *Physical Review E* 92 (4): 040802.

Cont, R. 2001. "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues." *Quantitative Finance* 1 (2): 223–236.

Das, S. R., S. Kim, and D. N. Ostrov. 2019. "Dynamic Systemic Risk: Networks in Data Science." *The Journal of Financial Data Science* 1 (1): 141–158.

De Spiegeleer, J., D. B. Madan, S. Reyners, and W. Schoutens. 2018. "Machine Learning for Quantitative Finance: Fast Derivative Pricing, Hedging and Fitting." *Quantitative Finance* 18 (10): 1635–1643.

Dempster, M. A. H., J. Kanniainen, J. Keane, and E. Vynckier. *High-Performance Computing in Finance: Problems, Methods, and Solutions*. Boca Raton: CRC Press, 2018.

Donoho, D. 2017. "50 Years of Data Science." *Journal of Computational and Graphical Statistics* 26 (4): 745–766.

Doshi-Velez, F., and B. Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." *arXiv* 1702.08608.

Easley, D., M. López de Prado, and M. O'Hara. 2016. "Discerning Information from Trade Data." *Journal of Financial Economics* 120 (2): 269–285.

Einav, L., and J. Levin. 2014. "The Data Revolution and Economic Analysis." *Innovation Policy and the Economy* 14 (1): 1–24.

Engle, R. F., and J. R. Russell. "Analysis of High Frequency Financial Data." In *Handbook of Financial Econometrics*, 2004.

Fabozzi, F. J. *Encyclopedia of Financial Models*, vol. 2. Hoboken: John Wiley & Sons, 2012.

Falkenberry, T. N. "High Frequency Data Filtering." White paper, Tick Data, Inc., 2002.

Fama, E. F., and K. R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–465.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. "From Data Mining to Knowledge Discovery in Databases." *AI Magazine* 17 (3): 37.

Flach, P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge, MA: Cambridge University Press, 2012.

Flannery, M. J., and K. W. Hankins. 2013. "Estimating Dynamic Panel Models in Corporate Finance." *Journal of Corporate Finance* 19: 1–19.

Fung, W., and D. A. Hsieh. 2000. "Performance Characteristics of Hedge Funds and Commodity Funds: Natural vs. Spurious Biases." *Journal of Financial and Quantitative Analysis* 35 (3): 291–307.

Giot, P. 2005. "Market Risk Models for Intraday Data." *The European Journal of Finance* 11 (4): 309–324.

Glasserman, P., and H. P. Young. 2016. "Contagion in Financial Networks." *Journal of Economic Literature* 54 (3): 779–831.

Goldstein, I., and L. Yang. 2017. "Information Disclosure in Financial Markets." *Annual Review of Financial Economics* 9: 101–125.

Guida, T. *Big Data and Machine Learning in Quantitative Investment*. Hoboken: Wiley Finance, 2018.

Haldane, A. G., and R. M. May. 2011. "Systemic Risk in Banking Ecosystems." *Nature* 469 (7330): 351.

Harvey, C. R. 2017. "Presidential Address: The Scientific Outlook in Financial Economics." *The Journal of Finance* 72 (4): 1399–1440.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. "… and the Cross-Section of Expected Returns." *The Review of Financial Studies* 29 (1): 5–68.

Hashem, I. A. T., I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan. 2015. "The Rise of 'Big Data' on Cloud Computing: Review and Open Research Issues." *Information Systems* 47: 98–115.

Heston, S. L., and N. R. Sinha. 2017. "News vs. Sentiment: Predicting Stock Returns from News Stories." *Financial Analysts Journal* 73 (3): 67–83.

Hey, A. J., S. Tansley, and K. M. Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*, vol. 1. Redmond, WA: Microsoft Research, 2009.

Holton, G. *Value-at-Risk: Theory and Practice*. Academic Press-Elsevier, 2003.

Huang, W., Y. Nakamori, and S. Y. Wang. 2005. "Forecasting Stock Market Movement Direction with Support Vector Machine." *Computers & Operations Research* 32 (10): 2513–2522.

Hurley, M., and J. Adebayo. 2016. "Credit Scoring in the Era of Big Data." *Yale JL & Tech*. 18: 148.

IEC. "Daily TAQ Client Specification." 2017. https://www.nyse.com/publicdocs/nyse/data/DailyT AQClient Specv3.0.pdf.

Ince, O. S., and R. B. Porter. 2006. "Individual Equity Return Data from Thomson Datastream: Handle with Care!" *Journal of Financial Research* 29 (4): 463–479.

Karppi, T., and K. Crawford. 2016. "Social Media, Financial Algorithms and the Hack Crash Theory." *Culture & Society* 33 (1): 73–92.

Kim, J. H., and P. I. Ji. 2015. "Significance Testing in Empirical Finance: A Critical Review and Assessment." *Journal of Empirical Finance* 34: 1–14.

Klaas, J. *Machine Learning for Finance: Principles and Practice for Financial Insiders*. Birmingham, UK: Packt Publishing, 2019.

Kofman, P., and I. G. Sharpe. 2003. "Using Multiple Imputation in the Analysis of Incomplete Observations in Finance." *Journal of Financial Econometrics* 1 (2): 216–249.

Kon, S. J. 1984. "Models of Stock Returns—A Comparison." *The Journal of Finance* 39 (1): 147–165.

Koop, G. *Analysis of Financial Data*. Hoboken: John Wiley & Sons, 2005.

Kovalerchuk, B., and E. Vityaev. "Data Mining for Financial Applications." In *Data Mining and Knowledge Discovery Handbook*, pp. 1203–1224. New York: Springer, 2005.

Kuhn, M., and K. Johnson. *Applied Predictive Modeling*, vol. 26. New York: Springer, 2013.

Lahmiri, S., and S. Bekiros. 2019. "Can Machine Learning Approaches Predict Corporate Bankruptcy? Evidence from a Qualitative Experimental Design." *Quantitative Finance* 19: 1569–1577.

Lo, A. W. 2017. "Moore's Law vs. Murphy's Law in the Financial System: Who's Winning?" *Journal of Investment Management* 15 (1): 17–38.

Lo, A. W., and A. C. MacKinlay. 1990. "Data-Snooping Biases in Tests of Financial Asset Pricing Models." *The Review of Financial Studies* 3 (3): 431–467.

López de Prado, M. L. 2015. "The Future of Empirical Finance." *The Journal of Portfolio Management* 41 (4): 140–144.

——. *Advances in Financial Machine Learning.* Hoboken: John Wiley & Sons, 2018.

——. 2019. "A Data Science Solution to the Multiple-Testing Crisis in Financial Research." *The Journal of Financial Data Science* 1 (1): 99–110.

Loughran, T., and B. McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-ks." *The Journal of Finance* 66 (1): 35–65.

Lyuu, Y. D. *Financial Engineering and Computation: Principles, Mathematics, Algorithms*. Cambridge, MA: Cambridge University Press, 2001.

Malkiel, B. G. 2003. "The Efficient Market Hypothesis and Its Critics." *Journal of Economic Perspectives* 17 (1): 59–82.

Mantegna, R. N. 1999. "Hierarchical Structure in Financial Markets." *The European Physical Journal B-Condensed Matter and Complex Systems* 11 (1): 193–197.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2019. "A Survey on Bias and Fairness in Machine Learning." *arXiv* 1908.09635.

Meinl, T., and E. W. Sun. "Methods of Denoising Financial Data." In *Handbook of Financial Econometrics and Statistics*, pp. 519–538. New York: Springer, 2015.

Mertzanis, C. 2018. "Complexity, Big Data and Financial Stability." *Quantitative Finance and Economics* 2: 637–660.

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267: 1–38.

Mills, T. C. 1995. "Modelling Skewness and Kurtosis in the London Stock Exchange FT-SE Index Return Distributions." *The Statistician* 44 (3): 323–332.

Mindell, J. 1961. "How News Affects Market Trends." *Financial Analysts Journal* 17 (1): 31–34.

Mitra, G., and L. Mitra. *The Handbook of News Analytics in Finance*, vol. 596. John Wiley & Sons, 2011.

Mitra, L., G. Mitra, and D. diBartolomeo. 2009. "Equity Portfolio Risk Estimation Using Market Information and Sentiment." *Quantitative Finance* 9 (8): 887–895.

Monk, A., M. Prins, and D. Rook. 2019. "Rethinking Alternative Data in Institutional Investment." *The Journal of Financial Data Science* 1 (1): 14–31.

Mullainathan, S., and J. Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Omur, U. *An Introduction to Computational Finance*, vol. 1. Singapore: World Scientific Publishing Company, 2008.

Petersen, M. A. 2009. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *The Review of Financial Studies* 22 (1): 435–480.

Rasekhschaffe, K. C., and R. C. Jones. 2019. "Machine Learning for Stock Selection." *Financial Analysts Journal* 75 (3): 70–88.

Rosenberg, B., and M. Houglet. 1974. "Error Rates in CRSP and COMPUSTAT Data Bases and Their Implications." *The Journal of Finance* 29 (4): 1303–1310.

Ruppert, D. *Statistics and Data Analysis for Financial Engineering*, vol. 13. New York: Springer, 2011.

Sewell, M. 2011. "Characterization of Financial Time Series." *RN* 11 (01): 01.

Seydel, R., and R. Seydel. *Tools for Computational Finance*, vol. 3. New York: Springer, 2006.

Simonian, J., and F. J. Fabozzi. 2019. "Triumph of the Empiricists: The Birth of Financial Data Science." *The Journal of Financial Data Science* 1 (1): 10–13.

Sirignano, J., and R. Cont. 2019. "Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning." *Quantitative Finance* 19 (9): 1449–1459.

Tanenbaum, A. S., and M. Van Steen. *Distributed Systems: Principles and Paradigms*. Upper Saddle River, NJ: Prentice-Hall, 2007.

Timmermann, A., and C. W. Granger. 2004. "Efficient Market Hypothesis and Forecasting." *International Journal of Forecasting* 20 (1): 15–27.

Tsay, R. S. *Analysis of Financial Time Series*, vol. 543. Hoboken: John Wiley & Sons, 2010.

Tu, J., and G. Zhou. 2004. "Data-Generating Process Uncertainty: What Difference Does It Make in Portfolio Decisions?" *Journal of Financial Economics* 72 (2): 385–421.

Van Liebergen, B. 2017. "Machine Learning: A Revolution in Risk Management and Compliance?" *Journal of Financial Transformation* 45: 60–67.

Varian, H. R. 2014a. "Beyond Big Data." *Business Economics* 49 (1): 27–31.

——. 2014b. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.

Verleysen, M., and D. Francois. "The Curse of Dimensionality in Data Mining and Time Series Prediction." In *International Work-Conference on Artificial Neural Networks*, pp. 758–770. New York: Springer, 2005.

White, G. I., A. C. Sondhi, and D. Fried. *The Analysis and Use of Financial Statements*. Hoboken: John Wiley & Sons, 1994.

Wolpert, D. H., and W. G. Macready. 1997. "No Free Lunch Theorems for Optimization." *IEEE Transactions on Evolutionary Computation* 1 (1): 67–82.

Wood, R. A., T. H. McInish, and J. K. Ord. 1985. "An Investigation of Transactions Data for NYSE Stocks." *The Journal of Finance* 40 (3): 723–739.

Zhang, D., and L. Zhou. 2004. "Discovering Golden Nuggets: Data Mining in Financial Application." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34 (4): 513–522.

Disclaimer
This article reflects the current opinions of the author, which are not necessarily those of Alphacruncher Hungary Kft.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

A Backtesting Protocol in the Era of Machine Learning

Rob Arnott, Campbell R. Harvey, and Harry Markowitz
The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/64

ABSTRACT: *Machine learning offers a set of powerful tools that holds considerable promise for investment management. As with most quantitative applications in finance, the danger of misapplying these techniques can lead to disappointment. One crucial limitation involves data availability. Many of machine learning's early successes originated in the physical and biological sciences, in which truly vast amounts of data are available. Machine learning applications often require far more data than are available in finance, which is of particular concern in longer-horizon investing. Hence, choosing the right applications before applying the tools is important. In addition, capital markets reflect the actions of people, who may be influenced by the actions of others and by the findings of past research. In many ways, the challenges that affect machine learning are merely a continuation of the long-standing issues researchers have always faced in quantitative finance. Although investors need to be cautious—indeed, more cautious than in past applications of quantitative methods—these new tools offer many potential applications in finance. In this article, the authors develop a research protocol that pertains both to the application of machine learning techniques and to quantitative finance in general.*

Triumph of the Empiricists: *The Birth of Financial Data Science*

Joseph Simonian and Frank J. Fabozzi
The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/10

ABSTRACT: *The authors situate financial data science within the broader history of econometrics and argue that its ascendance marks a reorientation of the field toward a more empirical and pragmatic stance. They also argue that owing to the unique nature of financial information, financial data science should be considered a field in its own right and not just an application of data science methods to finance.*

Rethinking Alternative Data in Institutional Investment

Ashby Monk, Marcel Prins, and Dane Rook

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/14

ABSTRACT: *As alternative data steadily become mainstream in finance, institutional investors may benefit from rethinking how they engage with alternative datasets. Specifically, they could gain from rethinking (1) alternative data's value proposition, (2) how they characterize alternative data, and (3) how they access alternative data. Rethinking their approaches to alternative data in these ways can help investors select alternative datasets that better align with their organizational resources and contexts. Such rethinking offers the greatest advantages when it focuses on building* defensive *and* defensible *strategies around alternative data, rather than prioritizing quicker exploitation of short-lived opportunities. Rethinking alternative data will require institutional investors to investigate new partnering possibilities, which should help them weather (and even thrive during) the escalating arms race among financial-market participants for alternative data. Building capacity for alternative data in these ways could also help investors accelerate innovation.*

PCA for Implied Volatility Surfaces

MARCO AVELLANEDA, BRIAN HEALY, ANDREW PAPANICOLAOU, AND GEORGE PAPANICOLAOU

MARCO AVELLANEDA
is a professor of mathematics in the Courant Institute of Mathematical Sciences at New York University in New York, NY.
avellaneda@courant.nyu.edu

BRIAN HEALY
is a researcher in the Department of Financial Computing and Analytics at University College London in London, UK.
brian@decisionsci.net

ANDREW PAPANICOLAOU
is an assistant professor in the Department of Finance and Risk Engineering at NYU Tandon School of Engineering in Brooklyn, NY.
ap1345@nyu.edu

GEORGE PAPANICOLAOU
is a professor in the Department of Mathematics at Stanford University in Stanford, CA.
papanicolaou@stanford.edu

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

KEY FINDINGS

- Principal component analysis of a comprehensive dataset of implied volatility surfaces from options on US equities shows that their collective behavior is captured by just nine factors, whereas the effective spatial dimension of the residuals is closer to 500 than to the nominal dimension of 28,000, revealing the large redundancy in the data.
- Portfolios of implied volatility surface returns, weighed suitably by open interest and Vega, track the principal eigenportfolio associated with a market portfolio of options, in analogy to equity portfolios.
- Retention of the tensor structure in the eigenportfolio analysis improves the tracking between the open interest–Vega weighted (tensor) implied volatility surface returns portfolio and the (tensor) eigenportfolio, indicating that data structure matters.

ABSTRACT: *Principal component analysis (PCA) is a useful tool when trying to construct factor models from historical asset returns. For the implied volatilities of US equities, there is a PCA-based model with a principal eigenportfolio whose return time series lies close to that of an overarching market factor. The authors show that this market factor is the index resulting from the daily compounding of a weighted average of implied-volatility returns, with weights based on the options' open interest and Vega. The authors also analyze the singular vectors derived from the tensor structure of the implied volatilities of S&P 500 constituents and find evidence indicating that some type of open interest- and Vega-weighted index should be one of at least two significant factors in this market.*

TOPICS: ***Statistical methods, simulations, big data/machine learning****

We show through principal component analysis (PCA) that a relatively small number of factors can account for most of the variation in the collective movements of the implied volatilities derived from US equity options. In fact, a matrix formed with normalized implied volatility returns over time has positive covariance with the first principal eigenvector, which is closely related to an open interest (OI)- and Vega-weighted basket of implied volatilities. This draws parallels with the first principal component obtained from the covariance of the matrix of normalized equity returns and its proximity to the capitalization-weighted portfolio (see Avellaneda and Lee 2010; Boyle 2014); OI is a measure of market size for options just as capitalization is a measure of market size for equities, and both implicitly

carry liquidity information. OI is the number of open contracts for a given option (name, strike, maturity) at a given time. Our findings highlight and give detailed insight into how PCA can be used to extract information from the covariance structure for a large dataset of implied volatilities and how new and improved implied volatility factors can be constructed using OI and Vega, both of which will be useful for portfolio and risk managers who have a need for better statistical prediction models to improve their estimated risk metrics (e.g., Value at Risk [VaR] and expected shortfall). To date, the preeminent volatility index is VIX, which is constructed from index options. The VIX has proven reliable but on occasion has shown susceptibility to outlier prices and manipulative trading (see Griffin and Shams 2017). The OI–Vega-based factors constructed in this article are more robust because they are based on hundreds of implied volatilities and place more emphasis on those contracts that have most trading interest.

The study described in this article builds on the work of Avellaneda and Dobi (2014), who considered a large dataset of implied volatility surfaces for a few thousand US equities and used PCA to find the smallest number of factors needed to explain the collective movements of these volatilities. They also constructed principal eigenportfolios and examined the qualitative structure of each from the first through fourth eigenvectors. We draw from the same data source as Avellaneda and Dobi (2014)—namely, the implied volatility surface (IVS) data available from OptionMetrics through Wharton Research Data Services. We hypothesize that the normalized covariance matrix of marketwide implied volatilities has a *low-rank plus random* structure (known as the *spike model*), and similar to Avellaneda and Dobi (2014), we find that removal of the low-rank components leaves a residual whose squared singular values are close in distribution to a Marchenko–Pastur (MP) law. Using random matrix theory (RMT), the presence of principal factors should make it possible to reject a model of purely random noise. RMT was used by Avellaneda and Dobi (2014), with the spectra-limiting MP distribution providing the basis for establishing cutoffs for the identification of the nonrandom structure. The analyses in this article consistently (for multiple years) show there to be at least two outliers in the singular value distribution, indicating that at least two factors drive the time series of IVS returns.

As noted, we focus on the first principal eigenportfolio and its closeness to various OI- and Vega-weighted portfolios. An eigenportfolio is a vector of portfolio weights that is derived from an eigenvector of the returns' covariance matrix. In equities, it is well known that the eigenportfolio constructed from the first principal eigenvector has explanatory power for the cross section of US equity returns (Avellaneda and Lee 2010; Boyle 2014) and that it tracks closely with a dominant factor such as a market portfolio. Perhaps the most significant finding in this article is a sizable body of evidence indicating that option OI and OI–Vega are key elements in the construction of factors to explain the collective cross-sectional changes of IVSs. Specifically, we show that various factors constructed from OI and OI–Vega weightings of implied volatility returns have significant explanatory power for interpreting the principal eigenportolio's returns. This finding can be considered the *implied volatility analogue* to the equity market's first principal factor, namely the capitalization-weighted returns portfolio. As noted, it appears that OI plays a role for implied volatility similar to that played by capitalization for equities. However, such a comparison is very informal because the capital asset pricing model (CAPM) and its related economic theory bind equities and capitalization closely together, whereas the implied volatility results shown in this article are, at present, statistical findings.

The time series of IVSs can be put into vector form, but its natural representation is a four-dimensional tensor, with the four dimensions being time, name, option maturity, and option delta (normalized strike). Our study of the first principal eigenportfolio can be extended to this tensor setting, which provides an example of how factor construction can, in fact, be improved by considering the natural representation offered by the tensor structure. The maturity and strike dimensions lend themselves to individual factors, for which we can construct individualized OI-weighted factors for each option maturity or for each maturity–delta pair. Individualized factors allow for a more nuanced weighing of changes in implied volatilities, and this leads to improved explanatory power for the covariance structure's, suitably defined, principal eigenportfolio.

REVIEW OF LITERATURE

The work of Avellaneda and Dobi (2014) and Dobi (2014) provides a cross-sectional classification of US

equity options based on implied volatility data for the period from August 2004 to August 2013, jointly with equity returns. The spectrum of the joint equity–IVS is used, in particular the leading eigenvalues, to classify options into those carrying mostly systemic risk and those carrying mostly idiosyncratic risk. Then, employing methods from PCA and results from random matrix theory, the significant eigenvalues are identified, and it is shown that approximately nine principal components suffice to reproduce the IVSs of all equities studied, with even fewer risk factors for so-called systemic names, such as SPY, QQQ, and AAPL. An explicit model is introduced that can be used to track the dynamics of the IVS, yet is compact and computationally tractable.

Focusing on the IVS for a single asset, Cont and Da Fonseca (2002) examined time series of option prices for options on the S&P 500 and FTSE 100 indexes. They showed how the IVS can be deformed and represented as a randomly fluctuating surface driven by a small number of orthogonal random factors. They found a simple factor model compatible with the empirical observations. Also of interest are methods for pricing baskets of many assets using option implied volatilities (see Avellaneda et al. 2002).

Before work on implied volatilities, there was a bounty of research on equities. The original work on portfolio composition dates back to Markowitz (1952); subsequent work on the CAPM model focused on the expected return of an asset relative to the risk-free instrument and derived a relationship between this and the excess return of a market or benchmark portfolio. Work in subsequent years suggested that factors other than the market excess return were being priced. In particular, Roll and Ross (1980) used data for individual equities during the 1962–1972 period and found that at least three are priced in the generating process of returns; Fama and French (1992) found the most significant factors to be market excess return, company size, and the ratio of the book value to the market value of the firm. Plerou et al. (2002) made early use of PCA and random matrix theory for analyzing equity returns.

The importance of eigenportfolios was highlighted by Boyle (2014), who examined the conditions under which frontier portfolios have positive weights on all assets. This is of interest because the market portfolio given by CAPM is mean–variance efficient and has positive weights on all assets. Before this work was that by Avellaneda and Lee (2010), who studied statistical arbitrage strategies in US equities with trading signals generated using PCA. Modeling the residuals of stock returns as a mean-reverting process, Avellaneda and Lee (2010) developed contrarian trading signals and then back tested these over the broad universe of US equities. The fact that these PCA-based strategies have an average annual Sharpe ratio that is statistically and economically significant provides empirical support for the PCA approach.

STRUCTURE AND RESULTS OF THE ARTICLE

This article has three main sections after this introduction. The first section addresses the estimation of the low-rank principal component structure from the standardized returns of options' implied volatility. The main result of this section is the introduction of an effective dimension approach for assessing the randomness of residuals, especially when there is only randomness in time because the vectorized IVS data produce residuals that do retain some of the structure of the data. The second section explores the role of the first principal component in constructing an eigenportfolio, with option OI and Vega as weights, as the primary factor in evaluating collective movements of IVSs. The final section makes use of the data's natural tensor structure for construction of improved principal eigenportfolios. The main contribution of this article is the presented evidence demonstrating the importance of OI when measuring changes in implied volatilities. Performing PCA to determine the number of relevant factors is a fairly standard procedure once we have standardized the data, but construction and analysis of eigenportfolios requires a deeper understanding of the data, including OI. The tensor analysis provides more depth of understanding because it shows us that construction of factors individualized to subcategories of options (e.g., separate OI-based factors for each of the options' maturities) leads to a clear improvement in the eigenportfolio's ability to account for implied volatilities' movements.

MATRIX OF IMPLIED VOLATILITY RETURNS

Let t be an index denoting calendar days. Let i be an index denoting an individual option contract, and denote the implied volatility for this particular option

contract as $\hat{\sigma}_i(t)$. We define at time t the vector of daily returns on the ith contract's implied volatility as

$$r_i(t) = \frac{d\hat{\sigma}_i(t)}{\hat{\sigma}_i(t)} \text{ for } 1 \le i \le N \text{ and } 1 \le t \le T, \tag{1}$$

where $d\hat{\sigma}_i(t) = \hat{\sigma}_i(t+dt) - \hat{\sigma}_i(t)$ with $dt = 1/252 = 1$ day. We assume throughout that the number of contracts far exceeds the number of days, $N \gg T$, which means that the covariance/correlation matrix has several eigenvalues equal to zero. We standardize these returns and then place them into a matrix $R \in \mathbb{R}^{N \times T}$, given by

$$R = \left[\frac{r_i(t) - \overline{r}_i}{h_i} \right]_{1 \le i \le N, 1 \le t \le T} \tag{2}$$

where $\overline{r}_i = \frac{1}{T} \Sigma_t r_i(t)$ and $h_i = \sqrt{\frac{1}{T-1} \Sigma_t (r_i(t) - \overline{r}_i)^2}$. Our hypothesis is that R can be decomposed into a low-rank factor matrix F and a random matrix X

$$R = F + X \tag{3}$$

which can be tested using the spike-model approach (see Benaych-Georges and Nadakuditi 2011). The low-rank matrix F can be further decomposed into orthogonal components

$$F = \sum_{i=1}^{d} f_i \theta_i^* \tag{4}$$

where each vector f_i is a principal characteristic of R with orthogonality between f_i and f_j $\forall i \neq j$, each θ_i is its loading, and $*$ denotes matrix/vector adjoint. In particular, if $||f_i|| = 1$, then $||\theta_i||^2$ is an eigenvalue of FF^*.

There are two obvious issues to address: the value of d in Equation 4 and the vectors $(f_i)_{i=1,\ldots,d}$. In the RMT literature it is equally as important to determine the θ_is because the criticality of a given θ_i will determine whether component f_i is distinguishable from R's bulk eigenvectors. In applications to financial data, however, the top components are typically much greater than the critical threshold, and attempts to include the middle ranks of F will usually lead to overfitting.

A brief description of the IVS data is given in the Appendix. For our analysis, we extracted 56 implied volatility values for each of the (roughly) 500 S&P 500 constituents for each of the approximately 252 business days in each year of data for 2012–2017. Specifically, we use options with time to maturity of 30, 60, 91, 122, 152, 182, 273, and 365 days and which have a delta (normalized strike; see Appendix) of −20, −30, −40, 50, 40, 30, and 20. Therefore, for each maturity we use three out-of-the-money put options (−20, −30, −40), one at-the-money option (50), and three out-of-the-money call options (40, 30, 20). Out-of-the-money options are used because these are more widely traded and hence are more liquid and have more reliable prices. Thus $N = 500 \times 8 \times 7 = 28{,}000$ and $T = 252$ (or 250 or 251 depending on when holidays fall in the year) if we use a one-year estimation window.

Singular Values of Nonprincipal Structure

A very basic estimator of the covariance matrix is $\hat{\rho} = RR^*/T$. Much of the literature has addressed methods of improving this estimator, including shrinkage of the eigenvalues by Ledoit and Wolf (2004) and asymptotic behavior of eigenvectors by Ledoit and Péché (2011) and Ledoit and Wolf (2012). Perhaps the most applicable reference for what we are seeking in this article is Benaych-Georges and Nadakuditi (2011), who gave a result stating that if the magnitude of the $||\theta||_i$'s is greater than some critical threshold with respect to the variance of X's entries in Equation 3, then the principal component vectors of $\hat{\rho}$ will be inside a cone centered around the principal eigenvectors of FF^*/T. For financial data, the first principal component often accounts for as much as 50% of the total variance and thus has an eigenvalue that is well over the threshold, but higher-order factors may be closer to the critical level. Detection of factors whose eigenvalues are near the critical threshold is interesting but not the main focus in this article. Instead, we focus on finding an estimate of the minimum number of factors needed to have a statistical nonrejection of the estimated low-rank model.

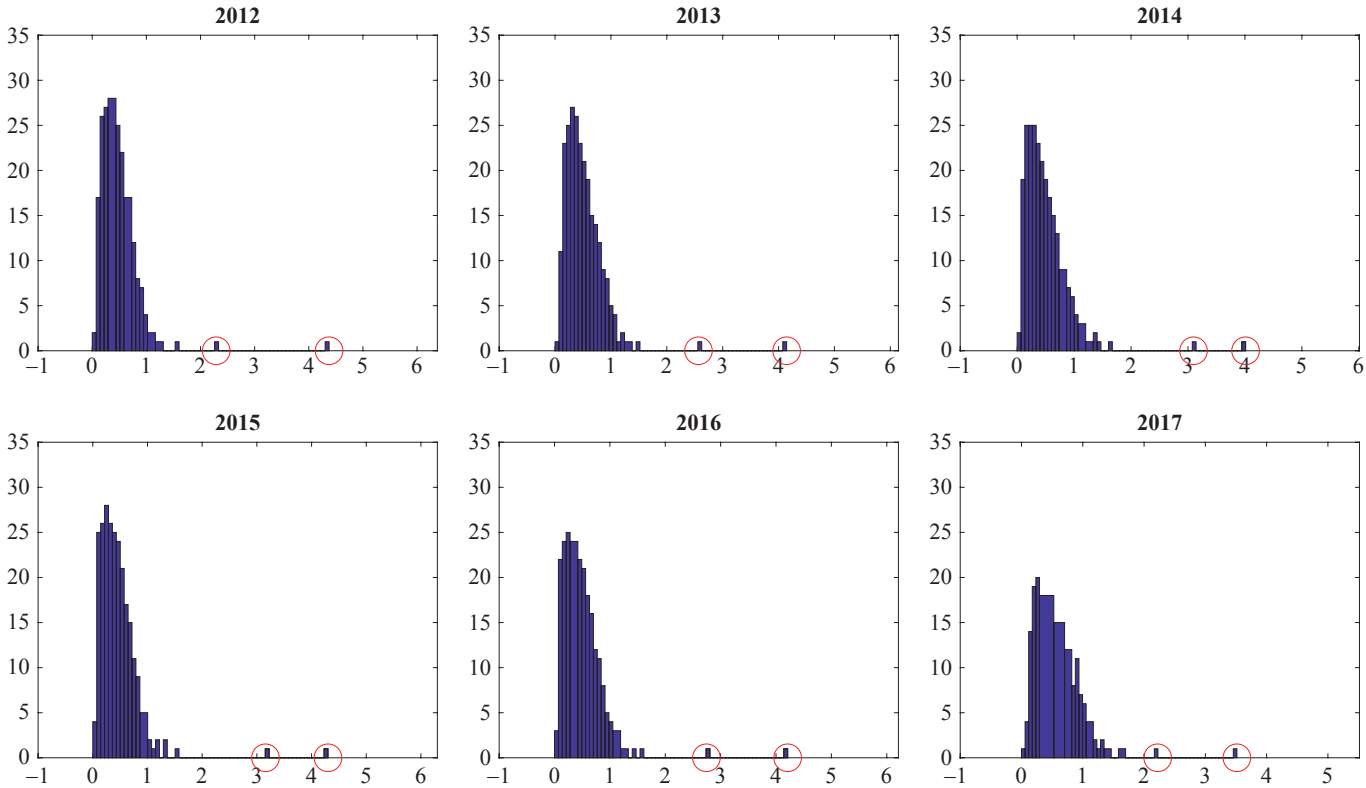
In practice, the estimator $\hat{\rho}$ is not calculated because usually $N \gg T$, making it more efficient to compute the singular value decomposition (SVD). The SVD represents R as

$$R = USV^*, \tag{5}$$

where $U = [U_1, U_2, \ldots, U_T]$ is an $N \times T$ matrix with orthonormal columns, S is a $T \times T$ diagonal matrix with entries $S_{11} \ge S_{22} \ge \ldots \ge S_{TT} \ge 0$, and $V = [V_1, V_2, \ldots V_T]$ is a $T \times T$ matrix with orthonormal columns.

EXHIBIT 1

Histograms of $\log(1 + S_{ii}^2/N)$ Where S_{ii} Are the Nonzero Singular Values of R, Estimated with 1-Year Windows from 2012 to 2017



Notes: If $F = 0$, there will be low probability of outliers from the bulk, and the histogram can be fitted with a Marchenko–Pastur density. However, in each of these histograms we see at least two outliers (circled marks), which indicates the rank of F is at least two. Hence, principal components need to be removed.

The nonzero eigenvalues of the correlation structure are the S_{ii}^2 values, and if R were completely random (i.e., if $F = 0$ in Equation 3), then the histogram of these values would be close to an MP density when N and T are large. However, from Exhibit 1, it is clear that at least two values are separated visibly from the bulk of $(S_{ii}^2/N)_{1\le i\le T}$. Hence, the rank of F is at least two, which means at least two principal components need to be removed from the data for the remainder to be considered noise.

For some $d \le T$, the best rank-d estimator of F that minimizes the Frobenius norm of error is

$$\hat{F} = \sum_{i\le d} S_{ii} U_i V_i^*,$$

with residual $\tilde{R} = R - \hat{F}$. The true rank of F is greater than d if a statistical test of the residual rejects the hypothesis of purely random entries in $\tilde{R}$. A Kolmogorov–Smirnoff (KS) test is likely to reject this hypothesis if d is too small. We devise a simple KS test using the data's own estimated MP distribution; that is, we use the estimated asymptotic distribution parameters obtained from $(S_{ii}^2/N)_{d<i\le T}$ and then check for significance of the associated KS statistic, as we now describe.

The MP density is

$$\nu(x) = \frac{1}{2\pi\gamma^2}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x} \quad \text{for } \lambda_- \le x \le \lambda_+ \qquad (6)$$

where

$$\lambda_\pm = \gamma^2(1 \pm \sqrt{\lambda})^2 \text{ and } \lambda > 0. \qquad (7)$$

If R were a random $N \times T$ matrix with independent identically distributed entries of mean zero and

variance γ^2, then for $N > T$ the RMT tells us that the empirical spectral distribution of the covariance $\frac{1}{N}R^*R$

$$\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\{S_{ii}^2/N\le x\}},$$

converges in probability (pointwise in x) to the distribution of the MP density in Equation 6 as N and T tend to infinity with $\lim\frac{T}{N}=\lambda\in(0,1)$ fixed. Conversely, the spectral distribution of $\frac{1}{T}RR^*$ is the case of $\lambda > 1$, wherein the limit law has an additional discrete mass at zero with weight $1-1/\lambda$, which appears because the $N\times N$ covariance has rank at most $T < N$, and so there are $N-T$ zero eigenvalues. Because we are interested only in eigenvalues through their empirical spectral density, we can consider the $T\times T$ covariance $\frac{1}{N}R^*R$ for which the dimension ratio $\lambda = T/N$ is less than one and there is no mass at zero in the asymptotic MP law.

The issue with the IVS data matrix R and its residual $\tilde{R}=R-\hat{F}$ is that we are not dealing with matrixes with independent identically distributed entries. We are, in fact, very far from it, and so it is not at all clear that the empirical spectral density of the residual matrixes will be close to the MP law. There are significant correlations among the entries of the residual matrixes, even without addressing the normalization issue. As noted, there is considerable theory on separating the bulk spectrum from the spike eigenvalues for idealized random matrix spike models; we may also cite the survey by Johnstone and Paul (2018) and, in dealing with normalization issues, El Karoui (2008). The theoretical criteria provided in the literature do not work with the IVS data, as expected. Writing a data matrix as a factor matrix plus a residual so that the residual is noise, or has no useful information, is a problem that arises often and in many different disciplines—not only with financial data but also, for example, in imaging in materials science (Berman 2019). With real data, this is almost always treated with a variety of empirical estimation methods whose validity is assessed on the basis of the results produced in specific applications.

For the IVS data, we will fit the empirical spectrum of the residuals to the MP law by matching supports, as we now describe. The quality of the fit is quantified by a KS test. The main result of this empirical fit, which works well for the IVS data, is to extract an estimated dimension ratio $\hat{\lambda}$ and standard deviation $\hat{\gamma}$ (Exhibit 2).

Exhibit 2

Estimated Parameters for the MP Distribution and Effective Dimension after Removal of Nine Principal Components for Each of the 251 or 250 IVS Daily Returns Observed in Each Year

Year	$\hat{\lambda}_+$	$\hat{\lambda}_-$	$\hat{\gamma}$	$\hat{\lambda}$	$\tilde{N}$
2012	1.77	0.07	0.80	0.44	573
2013	1.93	0.10	0.85	0.40	633
2014	2.15	0.07	0.86	0.49	519
2015	1.80	0.06	0.80	0.47	537
2016	1.98	0.06	0.83	0.50	506
2017	2.33	0.11	0.93	0.42	601

Notes: The data points are $X_i = S_{ii}^2/N$ for $i > 9$, the estimates are given by Equations 8 and 9, and the effective dimension is $\tilde{N} = T/\hat{\lambda}$. Note in particular that $\tilde{N} \ll N \sim 25{,}000$.

Let $X_i = S_{ii}^2/N$ be the eigenvalues (normalized) of the IVS data matrix R (normalized). Given the number of factors d that we want to retain, the estimators for the support $\lambda_\pm$ of the empirical spectral density, or histogram, of the residual that we use are

$$\hat{\lambda}_+ = X_{d+1} \text{ and } \hat{\lambda}_- = X_T, \tag{8}$$

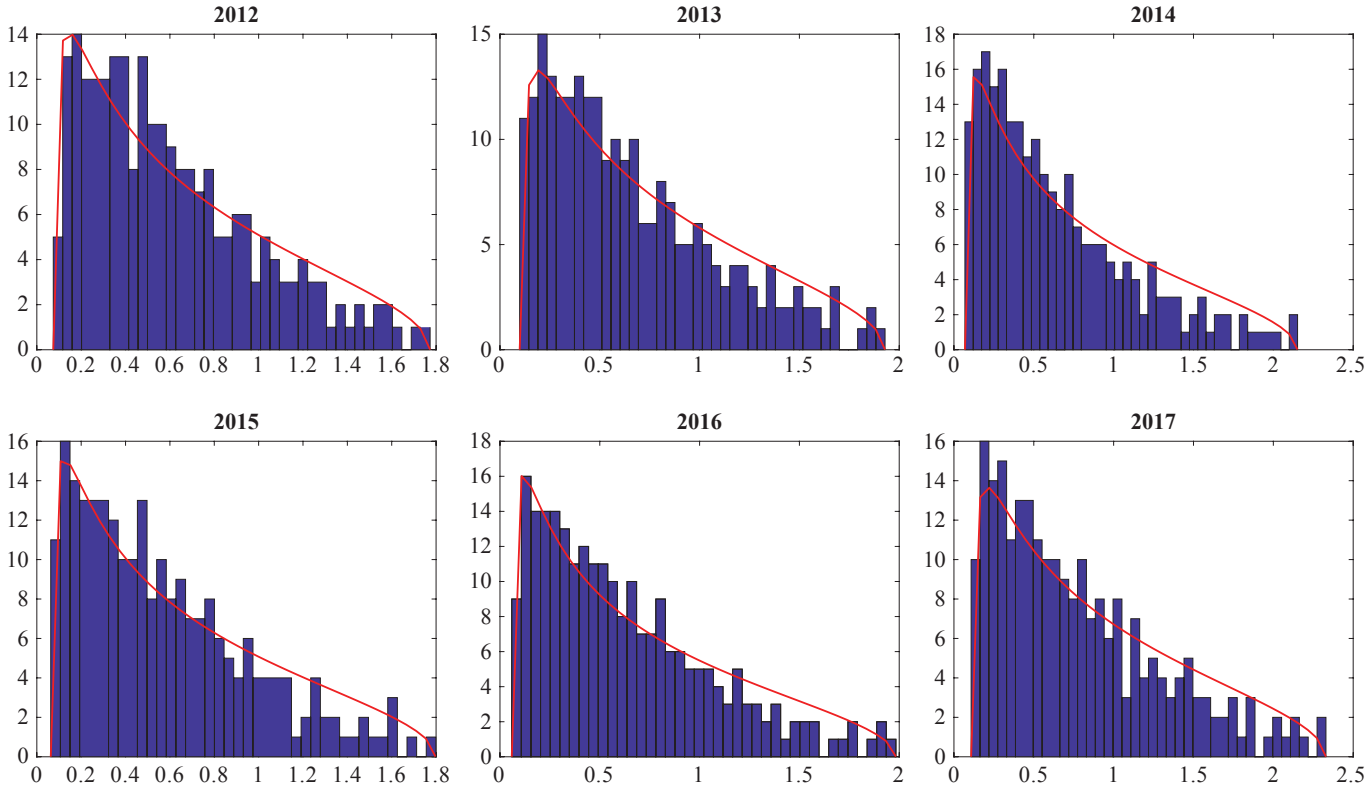
which, using Equation 7 with $\lambda\in(0,1)$, gives

$$\hat{\gamma}=\frac{\sqrt{\hat{\lambda}_+}+\sqrt{\hat{\lambda}_-}}{2} \text{ and } \hat{\lambda}=\left(\frac{\sqrt{\hat{\lambda}_+}-\sqrt{\hat{\lambda}_-}}{2\hat{\gamma}}\right)^2. \tag{9}$$

The fitted MP densities to the empirical spectral densities of the residuals with nine components removed are shown in Exhibit 3 for each of the years from 2012 to 2017. A two-sample KS test does not reject any of these years, and increasing d to 10, 20, 30, and 50 continues to result in a visibly good fit and nonrejection by the KS test. Hence, we conclude that nine factors is typically sufficient to describe the daily systematic movements among all IVSs over a single year of daily data. In contrast, for equity returns for the S&P 500 constituents, (roughly) 20 factors are typically required to account for the majority of daily movement, with the number dropping below 10 during the 2008 financial crisis (see Avellaneda and Lee 2010).

EXHIBIT 3

MP Density Fitted to Histogram of Remaining $T - d$ Squared Singular Values over N, 2012–2017



Notes: For each of the years from 2012 to 2017, we removed the top $d = 9$ principal singular values and fitted an MP density to the histogram of the remaining $T - d$ squared singular values over N (i.e., S_{ii}^2/N for $i > d$). A two-sample Kolmogorov–Smirnoff test does not reject any of these years. Removal of more than nine components also leads to nonrejection of the hypothesis that the residual matrixes $\tilde{R}$ have purely random entries.

A direct comparison with work by Avellaneda and Dobi (2014), wherein a cutoff for purely random entries was determined to be around 108 factors (out of ~3,000 names), is not possible because the dataset used by Avellaneda and Dobi (2014) includes equity returns normalized using the strike of the at-the-money option. This is a more complex dataset because the vectorized time series of IVS plus equities (of size 28,000 + 500 in the context of this article) is quite heterogeneous, and therefore the "low-rank plus random" decomposition needs additional attention.

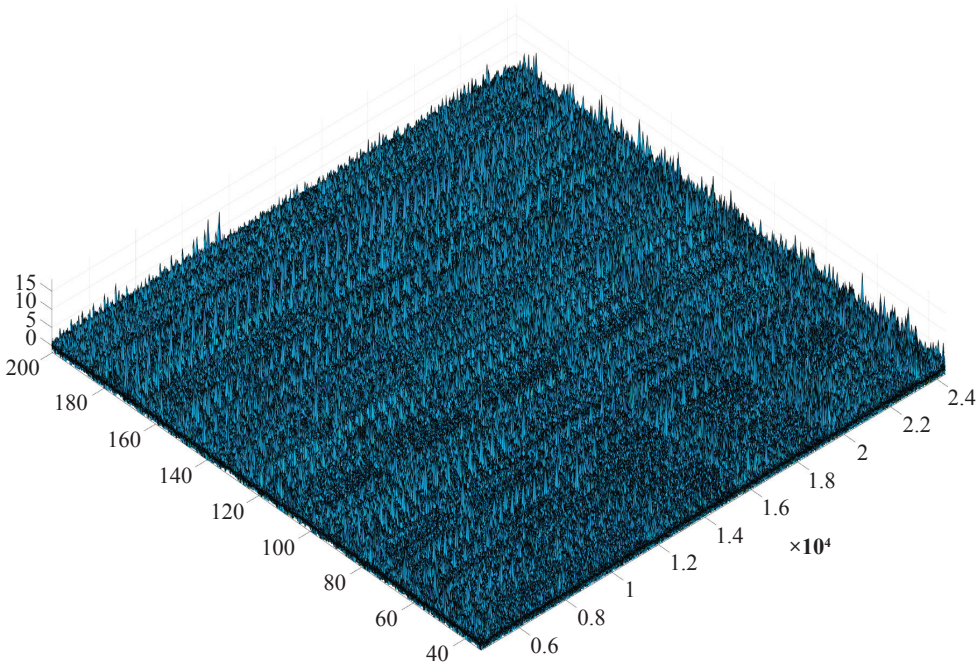
In this article, the nine factors (out of ~500 names for IVS) describe the systematic movements. Compared with equity returns by themselves, the number of factors is of comparable size for the different years between 2012 and 2017, and $d = 20$ factors are needed for the S&P 500 constituents to produce a random residual.

Effective Dimension in Residual Matrixes

We will now introduce the notion of effective dimension of the data as follows. The parameter estimates given in Equations 8 and 9 for the dimension ratio $\lambda = T/N$ and the data standard deviation γ imply that the empirical spectral density of the data fits an MP distribution associated with a random matrix whose dimensions are different from those of the actual data matrix R. This is because the IVS data has entries with significant correlations between them, and the effective dimension idea is a way to account for or quantify this feature. We could say roughly that the residual data matrix does not have independent entries, but it does have some kind of independence when the data are grouped in blocks and the number of such blocks plays the role of an effective dimension. This assessment is made using the empirical spectral density of the residual.

EXHIBIT 4

Nonrandom Structure Seen in the Residual Matrix $\tilde{R}$ Even after Sufficient Components Have Been Removed for Nonrejection by KS Test



Notes: The horizontal coordinate on the right is the IVS name, delta, and maturity (vectorized), and that on the left is time. This patterned structure suggests that the spatial modes are nonrandom and that the randomness we have concluded from the KS test is due to randomness in the temporal loadings. Random temporal loadings is precisely what is suggested in Equation 10.

Denoted by $\tilde{N} = T/\hat{\lambda}$, we call this the effective dimension associated with a residual that is purely random. In other words, the N-dimensional columns of U_{d+1}, U_{d+2}, … and U_T do not affect the spectrum, and we have

$$\frac{1}{N}\tilde{R}^*\tilde{R} = \frac{1}{N}\sum_{i=d+1}^{T} S_{ii}^2 V_i V_i^* \overset{\mathcal{D}}{\approx} \frac{1}{\tilde{N}} Y^* Y, \tag{10}$$

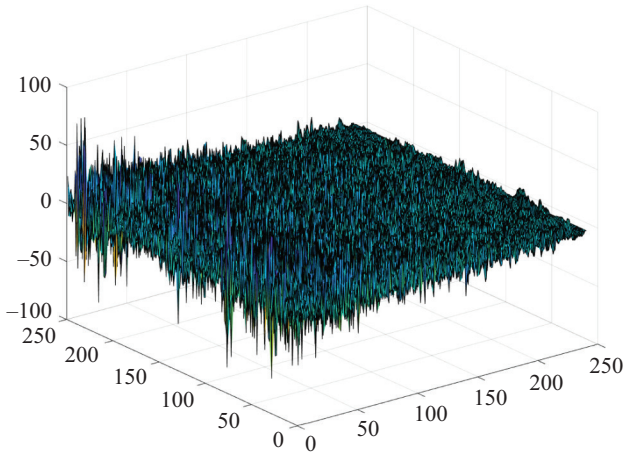
where Y is an $\tilde{N} \times T$ matrix with purely random entries and "$\overset{\mathcal{D}}{\approx}$" denotes approximate equality in distribution, in the sense of the fitting of the empirical spectral density to the MP law that was described in the previous section. Hence, we are looking for pure randomness in the temporal loadings, and we are not concerned if nonrandom structure remains in the higher-order spatial components. In fact, for the implied volatility data there are clear patterns of nonrandom structure in the residual $\tilde{R}$ even for d large enough for nonrejection of the randomness hypothesis, as is clearly seen in Exhibit 4 for the 2017 data. The reason for the nonrandom patterns is simple: The implied volatility's four-dimensional tensor structure was flattened into a two-dimensional $N \times T$ matrix using a lexicographical ordering for vectorizing the IVS data (name, strike, and maturity) that prevails even after removal of PCA factors. This patterned structure does not mean we have incorrectly concluded there is pure randomness in the residual as seen by the empirical spectral density; rather, it suggests that the spatial modes (name, strike, and maturity) retain some of their structure and that the randomness we have found from the KS test is due to randomness in the temporal loadings. Indeed, randomness of temporal loadings is precisely what is suggested by the approximate distributional equivalence expressed in Equation 10.

Denote the tth column of R as R_t, which we can write as

$$R_t = \sum_{i=1}^{d} f_i \theta_{it} + \sum_{i=d+1}^{T} S_{ii} U_i V_{it},$$

EXHIBIT 5

Matrix $S\tilde{V}^*$ of Temporal Loadings for the 2017 Data of Dimension $(T - d) \times T$ with $d = 9$



Note: The distinctive tapering is not an indication of nonrandomness in the residual but instead says that $S\tilde{V}^*/N$ is close in distribution to ΣQ^*, where $Q\Sigma Q^*$ is the spectral form of another matrix $Y^*Y/\tilde{N}$ and Y is an $\tilde{N} \times T$ matrix with purely random entries of variance $\hat{\gamma}^2$.

where θ_{it} and V_{it} are the tth entry of θ_i and V_i, respectively; U_i is the ith column of U. Clearly θ_{it} are the temporal loadings on the ith principal factor, and for $i > d$ the temporal loadings are $S_{ii}V_{it}$. Denote the higher-order temporal modes as $\tilde{V} = [V_{d+1}, \ldots, V_T]$. The KS test has indicated randomness of these loadings for $i > d$, which means the $(T - d) \times T$ matrix $S\tilde{V}^*$ has a covariance spectrum close in distribution to that of a random matrix. However, this matrix has a distinctive tapering that would counter any claim of pure randomness. Indeed, Exhibit 5 shows this tapering for the 2017 data for S_{ii}^2/N with $i > 9$ (i.e., with $d = 9$). However, if one considers a purely random matrix Y of dimension $T \times \tilde{N}$ with entries having variance $\hat{\gamma}^2$ and expresses it in spectral form such that $Y^*Y/\tilde{N} = Q\Sigma Q^*$, then ΣQ^* is a tapered matrix that is derived from a purely random matrix. Moreover, the squared singular values of ΣQ^* fit the MP law and are equal in distribution to the squared singular values of $S\tilde{V}^*/N$.

The notion of effective dimension as we have introduced it here is robust and useful in understanding the real information content of the IVS data. Note in particular that in Exhibit 2, which has the parameter estimators for the MP fit as well as the effective dimension for each of the 6 years we have considered, a striking finding is that we have a very low effective dimension for each year (i.e., $500 \sim \tilde{N} \ll N \sim 25{,}000$). The effective dimension of the IVS residuals in a factor decomposition indicates that there is structure in these residuals; they are not purely random matrixes.

PRINCIPAL EIGENPORTFOLIOS AND OI-WEIGHTED INDEXES

The portfolio constructed from the first eigenvector of the normalized covariance, the eigenportfolio, has been analyzed by Avellaneda and Lee (2010) and Boyle (2014) and others for equities returns. In this section, we will apply a similar approach and terminology to construct a portfolio of implied volatilities and compare it with an analog of a market portfolio. Although such portfolios can exist in theory, they are not directly[1] tradeable. However, it is still of interest to introduce them and use them as a proxy for the market's collective volatility, similar to the VIX index.

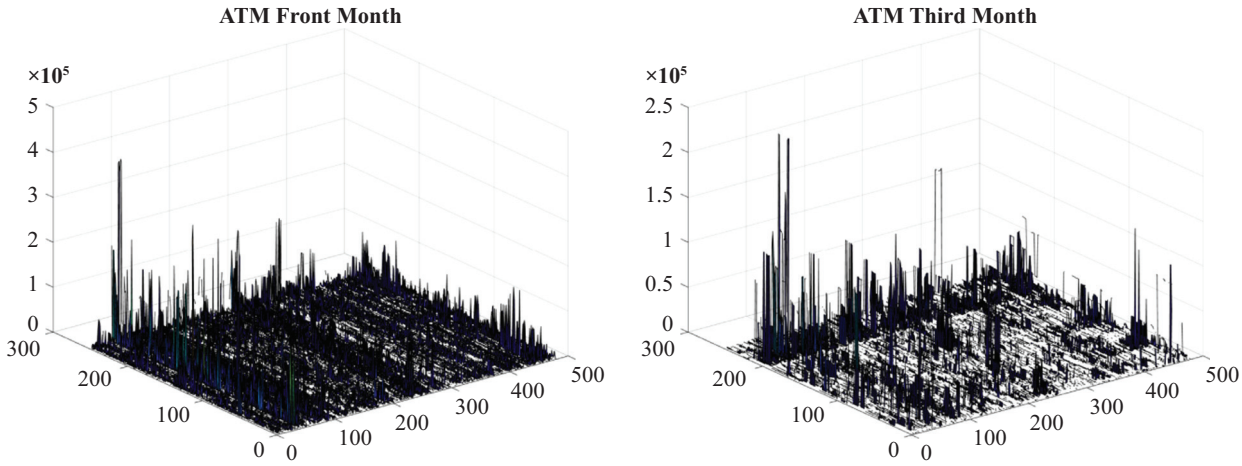
Central to our analysis of the eigenportfolio is its closeness to an OI-weighted factor of our construction. Analogous to the case in equities wherein the eigenportfolio's returns track the capitalization-weighted market portfolio (with varying degrees of closeness), we show that the implied-volatility eigenportfolio's returns track closely to various OI-weighted factors or indexes. It is essential to include the OI because it is the implied volatilities' analogue to capitalization, thereby capturing each option contract's importance.[2] In the analysis of both implied volatilities and equities, the underlying spike model (see Benaych-Georges and Nadakuditi 2011) provides a mathematical framework for the closeness of the eigenportfolio and factor returns. Exhibit 6 shows daily OI amounts for at-the-money options for 2017. It is important to note the spikes and the zeros in the OI. Most options have zero open trades at any point in time, which is an indication that a rather small percentage of options can explain changes in implied volatilities.

[1] A portfolio can be constructed with returns equal to a linear combination of implied volatilities returns (i.e., a linear combination of the $r_i(t)$ values). However, such a portfolio would probably incur a roll yield owing to the daily rebalancing required to maintain a constant maturity and constant delta position.

[2] In practice, the CBOE's VIX calculation uses OI as well. Specifically, the VIX construction includes only SPX options that are between the at-the-money mark and the last strike before the first two consecutive strikes that have zero OI.

EXHIBIT 6

Daily Values of OI for the At-the-Money Options with 30 Days to Maturity (front month) and 90 Days to Maturity (third month)



Notes: The horizontal coordinate on the left is time and that on the right is option name. Note the spikes and the zeros in the OI, which indicate that only a small percentage of options can explain changes in implied volatilities because most options do not have open trades any on given day.

Trading-Implied Volatility and Construction of OI-Based Factors

It is not only OI, however, that must be taken into consideration in constructing IVS returns portfolios. We must also account for the sensitivity of IVS to fluctuations as measured by the Vega, for example. We can motivate this sensitivity in portfolios we are about to create by first introducing a synthetic market of exchange-traded notes (ETNs). For each option we can think of an ETN whose prospectus states the daily returns to be

$$\frac{dE_i(t)}{E_i(t)} := r_i(t), \tag{11}$$

where $r_i(t)$s are the components of the implied volatility return vector defined in Equation 1 and $dt = 1/252 = 1$ day. The ETN whose returns are given by Equation 11 is the stochastic component in the returns of a Δ-neutral options position. Indeed, letting $C_i(t)$ denote the option price with underlying price, time to maturity, and delta $(S_i(t), \tau_i, \Delta_i)$, from Itô's lemma we compute the (unitless) differential of a Δ-neutral position (up to a term of size Big-Oh),

$$\begin{aligned}\frac{dC_i(t) - \Delta_i dS_i(t)}{S_i(t)} &= \frac{\mathcal{V}_i(t)}{S_i(t)} d\hat{\sigma}_i(t) + O(dt) \\ &= \mathcal{V}_i^{untls}(t) r_i(t) + O(dt),\end{aligned} \tag{12}$$

where $\mathcal{V}_i(t)$ is the Vega for the ith option and

$$\mathcal{V}_i^{untls}(t) = \frac{\hat{\sigma}_i(t)\mathcal{V}_i(t)}{S_i(t)}, \tag{13}$$

is a unitless Vega, which is the dollar-Vega divided by the price of the underlying. We will use Equation 13 in constructing weighing factors, in addition to OI, after we first discuss them more generally.

We would like to construct a global factor that can describe upward of 50% of daily variance for all of the IVS ETNs. In equities, such a factor is the market portfolio, which suggests to us that the number of outstanding shares (contracts) should have some bearing on the relevance of an individual equity in factor construction. Indeed, in options this is precisely the OI, and a general form for a global factor with only OI weighing is

$$\frac{dQ(t)}{Q(t)} := \frac{\sum_i \omega(\mathcal{OI}_i(t)) r_i(t)}{\sum_i \omega(\mathcal{OI}_i(t))}, \tag{14}$$

where $\mathcal{OI}_i(t)$ is the OI for the ith option and $\omega(\cdot)$ is a weighting function of our choosing. In the simplest case we have $d = 1$ in Equation 4, and the ETN returns have a simple factor-based returns model

$$\frac{dE_i(t)}{E_i(t)} = \beta_i \frac{dQ(t)}{Q(t)} + \xi_i(t), \tag{15}$$

where ξ_i is an idiosyncratic noise component independent of $Q(t)$. Ordinary least squares regression shows us that the β_is are given by the covariance with the factor

$$\beta_i = \mathrm{cov}\left(\frac{dE_i}{E_i}, \frac{dQ}{Q}\right) / h_q^2, \tag{16}$$

where $h_q^2 = \mathrm{var}(\frac{dQ}{Q})$

Remark 1. Boyle (2014) explained how the principal eigenportfolio is a frontier portfolio if the Perron–Frobenius theorem applies. However, for options, the frontier/CAPM theory does not apply because the lifetime of an option is too short. Therefore, comparisons with equities are merely an informal, statistical analogy.

The Spike Model and the Principal Eigenportfolio

Let us consider the $N \times N$ empirical covariance matrix for the returns of the synthetic ETNs

$$\hat{\Sigma}_{ij} := \frac{1}{T-1} \sum_{t=1}^{T} (r_i(t) - \overline{r}_i)(r_j(t) - \overline{r}_j) \quad \text{for} \quad 1 \le i, j \le N.$$

Using the returns model in Equation 15, the population covariance matrix is

$$\Sigma = h_q^2 \beta\beta^* + \Omega, \tag{17}$$

where Ω is the covariance matrix with $\Omega_{ij} = cov(\xi_i(\cdot), \xi_j(\cdot))$, and $h_q^2 = \mathrm{var}\left(\frac{dQ(\cdot)}{Q(\cdot)}\right)$. Equation 17 is a spike model, as referred to earlier, because the βs describe a substantial portion of variance and cause a single eigenvalue to stick out from the rest of the spectrum. The first principal component of the ETN empirical covariance matrix will be nearly proportional to the βs of the OI-weighted portfolio if the $\xi(t)$ covariance is not too large. For the model in Equation 3, the distribution of the empirical matrix principal component is shown by Benaych-Georges and Nadakuditi (2011) to be within a cone surrounding the spike model's low-rank component if the difference between $||\theta||^2$ and the variances of the noise is over a critical amount. For the spike model in Equation 17, the critical threshold is crossed if $h_q^2 \, ||\beta||^2$ exceeds a threshold determined by the covariances of the $\xi_i(t)$'s, which should happen as N grows.

In finance, there usually are differing sizes among the Σ_{ii}s, which means better statistical estimation of principal eigenvectors results from consideration of correlations rather than covariances; this is the normalization issue we noted earlier. The ETNs' empirical correlation matrix is

$$\hat{\rho} = h^{-1} \hat{\Sigma} h^{-1},$$

where h is a diagonal matrix of standard deviations h_i defined in Equation 2. Letting u_1 denote the principal eigenvector of $\hat{\rho}$, the spike model suggests $u_1 \approx \frac{1}{c} h^{-1} \beta$ for c, a normalizing constant. Using another (orthogonal) eigenvector $\tilde{u}$ such that $\tilde{u} \perp u_1$, we can also construct portfolios as done by Avellaneda and Lee (2010):

$$\pi_1 = \frac{h^{-1} u_1}{\sum_i (h^{-1} u_1)_i}, \quad \tilde{\pi} = \frac{h^{-1} \tilde{u}}{\sum_i (h^{-1} \tilde{u})_i},$$

which are orthogonal in the sense that covariance of these portfolios' returns is zero

$$\pi_1^* \hat{\Sigma} \tilde{\pi} = u_1^* \hat{\rho} \tilde{u} = \lambda_1 u_1^* \tilde{u} = 0,$$

where $\lambda_1 > 0$ is the principal eigenvector such that $\hat{\rho} u_1 = \lambda_1 u_1$.

Proposition 0.1. Returns of the top eigenportfolio tend toward the factor returns plus some tracking error. Returns of the orthogonal portfolios tend toward factor neutrality.

Proof. Assuming the parameters are such that we are over the critical levels noted by Benaych-Georges and Nadakuditi (2011) and letting $EP_1(t)$ be the principal eigenportfolio, we have

$$\begin{aligned} &\frac{dEP_1(t)}{EP_1(t)} - \frac{\sum_i (\beta_i / h_i)^2}{\sum_i \beta_i / h_i^2} \frac{dQ(t)}{Q(t)} \\ &= \left(\sum_i \left(\pi_{1i} - \frac{\beta_i / h_i^2}{\sum_j \beta_j / h_j^2} \right) \beta_i \right) \frac{dQ(t)}{Q(t)} + \sum_i \pi_{1i} \xi_i(t) \\ &\to \varepsilon_1(t), \end{aligned}$$

where factor returns disappear because $\pi_{1i} - \frac{\beta_i / h_i^2}{\Sigma_j \beta_j / h_j^2} \to 0$ as the random matrix's dimensions grow, and where

$\varepsilon_1(t) = \lim_N \Sigma_i \pi_{1i} \xi_i(t)$ is the tracking error. All orthogonal portfolios $\tilde{\pi}$ are approximately factor neutral:

$$\frac{d\widetilde{EP}(t)}{\widetilde{EP}(t)} = \underbrace{\left(\sum_i \tilde{\pi}_i \beta_i\right)}_{\approx 0} \frac{dQ(t)}{Q(t)} + \sum_i \tilde{\pi}_i \xi_i(t) \to \tilde{\varepsilon}(t),$$

where $\tilde{\varepsilon}(t) = \lim_N \Sigma_i \tilde{\pi}_i \xi_i(t)$, and where the limit happens because $(h^{-1}\tilde{u})^* \beta \approx \tilde{u}^* h^{-1} h u_1 = 0$ becomes more accurate and tends toward an equality as the gap between $||h^{-1}\beta||$ and the critical level increases. □

Empirical Analysis

The first spatial[3] singular vector U_1 computed by the SVD in Equation 5 is the empirical estimator for u_1, and the estimator for each h_i is the empirical standard deviation of each r_i. Hence, we can compute the eigenportfolio from the data and then compare it with the empirically estimated βs. In our studies we will consider two weighting functions

$$\omega(\mathcal{OI}) = \mathcal{OI} \quad \text{or} \quad \omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}, \quad (18)$$

where $\mathcal{V}^{untls}$ denotes the unitless Vega of Equation 13. Generally speaking, $\omega(\mathcal{OI}) = \mathcal{OI}$ results in a factor with a less significant intercept in ex post regressions of eigenportfolio returns onto the factor returns, whereas $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$ results in the same regression having a significant intercept but lower projection error. The plain OI weighting is a bit strange, however, because it counts contracts without taking into account the sensitivity of the contract to a change in the volatility of the underlying stock. It should also be noted that the log weighting has a factor loading that is closer to unity, whereas the plain OI weighting is significantly less than unity.

The OI and unitless Vega weighing does, in fact, matter, and it performs better when the tensor data structure is taken into consideration. We have chosen the combined OI and unitless Vega weighing in the form $\log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$, which works well for the IVS data, although it is rather arbitrary at present.

In Exhibit 7, we see the comparison for each of the years, with each year's eigenportfolio computed using the 251 or 250 days of data from the year and with a $Q(t)$ factor computed using the weight function $\omega(\mathcal{OI}) = \mathcal{OI}$. If the eigenportfolio's weights are sorted in descending order (i.e., we sort π_1 in descending order) and the sorting index then is used to permute the vector $\text{diag}^{-2}(h)\beta / \Sigma_i (\text{diag}^{-2}(h)\beta)_i$, then the sorted vector and the permuted vector should line up. Indeed, the plots in Exhibit 7 show this lining up, with the sorted eigenportfolio in red and the permuted βs in blue.

We can also check the name, the tenor, the Δ, the β_i / h_i^2's, and the OI for the top-weighted options. These traits are listed in Exhibit 8 for the top 32 options in the 2012 sorting. It is interesting to note that most of the top options are out-of-the-money put options, all with 365 days to maturity (the longest-dated options in the dataset). Generally speaking, long-dated options have higher Vega[4] and therefore are most sensitive to changes in implied volatility. The interpretation of long-dated options dominating the first eigenportfolio is simple: The first eigenportfolio explains the most systematic movements among the options and should be the least sensitive to idiosyncratic noise; it therefore ignores short-dated options that may fluctuate idiosyncratically because of short-lived risk events. The years 2013–2017 had similar characteristics in the top 32 eigenportfolio options.

We also perform checks to make sure eigenportfolios can track in an online setting; that is, all quantities used to compute a portfolio in real time are adapted to the filtration generated by the options data (Exhibit 9).[5] We group all 6 years of data into one simulated run and then compute an eigenportfolio each month using the previous 6 months' daily returns. The returns are compounded daily, but we only update the portfolio weights monthly. Exhibit 9, Panels A–D, show the

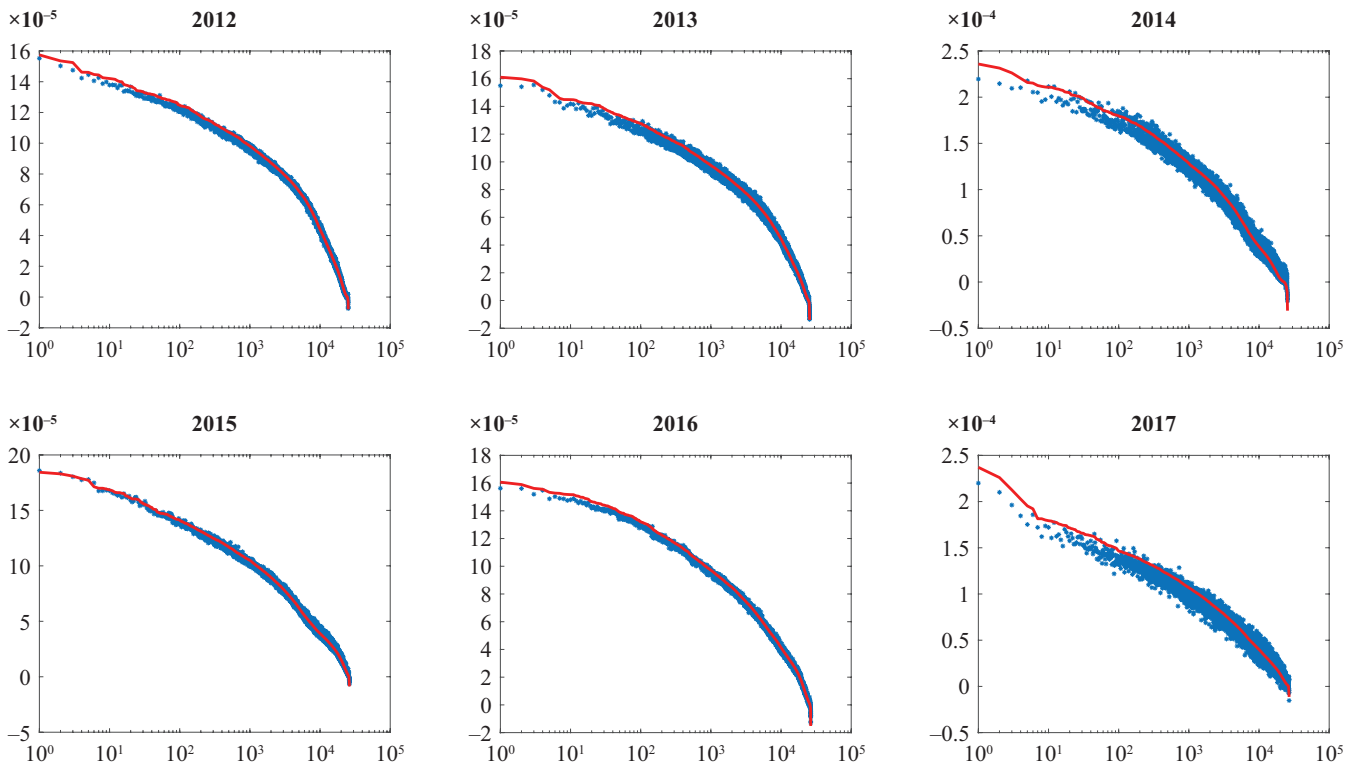
[3] That is, of the IVS vector of names, deltas, and maturities.

[4] The Black–Scholes call/put option Vega is $\mathcal{V}(t) = SN'(d_1)\sqrt{\tau}$, where τ is the time to maturity. Hence, all other things being equal, longer-dated options have higher Vega.

[5] The online eigenportfolio that we compute has one major anticipatory element: survivorship bias. We selected names of the S&P 500 constituents from 2017 and collected their options' data going back to 2012. Hence, there is some survivorship bias in favor of names that performed well enough to stay in the index for these 6 years. We acknowledge this bias and realize that it will persist over time, but for our analysis of daily statistics it does not make our findings any less valid. Moreover, the OI weighing favors liquid options for which survivorship bias is likely to be reduced.

EXHIBIT 7

Eigenportfolio Computed from the Data (red) Compared with the Theoretical Eigenportfolio That Is Close to β_i/h_i^2 (blue)



Notes: The horizontal axes are in log scale. Each $\beta_i = \mathrm{cov}(dE_i(\cdot)/E_i(\cdot), dQ(\cdot)/Q(\cdot))/h_q^2$, *where for these plots the factor Q(t) has been computed according to the formula in Equation 14 using the weighting function* $\omega(\mathcal{OI}) = \mathcal{OI}$. *To generate this plot, we first sort the eigenportfolio in descending order and then insert the sorting index into the vector* $h^{-2}\beta/\sum_i(h^{-2}\beta)_i$. *The lining up of the two vectors using a single sorting index is evidence that the factor computed using the weighting function* $\omega(\mathcal{OI}) = \mathcal{OI}$ *is close to the data's principal component.*

online six-month sliding window performance of the eigenportfolio alongside the factor returns and the VIX returns. Note in Exhibit 9, Panel B, that the eigenportfolio returns have significant explanatory power for the returns on VIX (i.e., $d\mathrm{VIX}(t)/\mathrm{VIX}(t)$) but also that returns for the OI factor and eigenportfolio have a much stronger linear dependence. Indeed, even if we use the weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$, which allows the OI factor to have better tracking with the VIX (see Exhibit 9, Panel D), the stronger linear dependence in daily returns between factor and eigenportfolio still prevails.

Finally, some discussion on performance with the two different weighting functions is in order. The visual evidence in Exhibits 7 and 9 should sufficiently convince the reader that OI needs to be included in factor construction. The discussion that remains is for us to decide how to determine the OI-weighted factor that is somehow best. Exhibit 10 provides evidence indicating that $\omega(\mathcal{OI}) = \mathcal{OI}$ is good because it leaves the least amount of unexplained systematic return after regression of the eigenportfolio onto the factor's returns (as well as two other factors). However, visually from Exhibit 9, Panel D, we see that $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$ produces a factor that is better for tracking the VIX; the VIX is the US market's premier volatility index, but this choice of ω leaves a significant amount of unexplained systematic return after the regression. Bear in mind that leaving unexplained systematic return is not entirely bad because we have thus far only been able to construct a single factor for explained implied volatility movements, but we know that we need at least two (recall Exhibit 1,

EXHIBIT 8

Ordering of Names in Top 32 Slots in Eigenportfolio of S&P 500 Constituents' Implied Volatility for 2012

Ticker	Maturity (days)	Δ	β/h^2	OI (average)
ADSK	365	–20	3.5332×10^3	2,751
KLAC	365	–20	3.4206×10^3	3,962
DHR	365	–20	3.3575×10^3	18,017
KLAC	365	–30	3.2402×10^3	3,906
LRCX	365	–30	3.2904×10^3	4,063
KLAC	365	–40	3.1996×10^3	5,171
LRCX	365	–40	3.2450×10^3	4,002
INTU	365	–40	3.1577×10^3	4,229
INTU	365	–30	3.1832×10^3	5,457
ADI	365	–20	3.1372×10^3	15,907
XLNX	365	–20	3.1402×10^3	4,237
FLR	365	–20	3.1365×10^3	19,285
ADI	365	–30	3.0942×10^3	13,276
ADSK	365	–30	3.1348×10^3	4,018
TXN	365	–20	3.1324×10^3	19,669
DHR	365	–30	3.0415×10^3	23,395
CHRW	365	–40	3.0531×10^3	2,874
CHRW	365	–30	3.0543×10^3	3,044
FLR	365	–30	3.0444×10^3	22,425
ROST	365	–30	3.1108×10^3	5,809
RCL	365	–40	3.0327×10^3	7,943
FDX	365	–30	3.0153×10^3	34,131
RCL	365	–20	3.0019×10^3	24,656
EXPD	365	–20	2.9784×10^3	5,796
LRCX	365	40	3.0237×10^3	14,541
FDX	365	–20	2.9778×10^3	43,925
FLR	365	40	2.9695×10^3	18,146
LRCX	365	30	3.0293×10^3	12,629
ADSK	365	–40	2.9768×10^3	3,314
SBUX	365	–40	3.0220×10^3	39,345
IR	365	–20	2.9369×10^3	4,699
CF	365	–20	2.9476×10^3	72,143

Notes: The maturity for all these top-weighted options is 365 days, the longest-dated options in our dataset. Generally, long-dated options have higher Vega, which means that the first eigenportfolio finds the most systematic explanation for implied volatility surface movements by considering the options with the greatest sensitivity to implied volatility changes.

where it was clear that there are at least two principal components).

It turns out that there is a setting wherein $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$ is an acceptable factor, but it will require us to move from ordinary flat linear algebra and use tensors in the next section.

FACTORS AND EIGENPORTFOLIOS USING TENSORS

The IVSs have a natural tensor structure with four dimensions: time, name, maturity, and delta (normalized strike). To work with this tensor, we first need to redefine our notation for implied volatility returns from the definition given in Equations 1 and 2. Before we had $1 \leq i \leq N$, where $N = 500 \cdot 8 \cdot 7 = 28{,}000$ (500 names, eight maturities, seven deltas). Now we have multiple indexes:

$$r_{ijk}(t) = \text{time-}t \text{ implied volatility return}$$
$$\text{for } i\text{th name, } j\text{th maturity and } k\text{th } \Delta,$$

where $1 \leq i \leq N^{(1)}$, $1 \leq j \leq N^{(2)}$, $1 \leq k \leq N^{(3)}$, and $1 \leq t \leq T$ (i.e., $N^{(1)} = 500$, $N^{(2)} = 8$ and $N^{(3)} = 7$). We standardize the returns and place them in a four-dimensional tensor

$$R = \left[\frac{r_{ijk}(t) - \overline{r}_{ijk}}{h_{ijk}} \right]_{1 \leq t \leq T,\ 1 \leq i \leq N^{(1)},\ 1 \leq j \leq N^{(2)},\ 1 \leq k \leq N^{(3)}}, \tag{19}$$

where $\overline{r}_{ijk} = \frac{1}{T}\Sigma_t r_{ijk}(t)$ and $h_{ijk} = \sqrt{\frac{1}{T-1}\Sigma_t (r_{ijk}(t) - \overline{r}_{ijk})^2}$. Similar to the sector-based hierarchical PCA done for equities by Avellaneda (2019), we can define individualized factors for each value of a certain tensor dimension. The maturity dimension and the delta dimension are the two candidates for individualized factors; the following subsection demonstrates the advantage of constructing the individualized factors.

Eigenportfolios via Multilinear SVD

A tensor analogue for the SVD is the multilinear singular value decomposition (MLSVD) (see De Lathauwer, De Moor, and Vandewalle 2000; Kolda and Bader 2009; Cichocki et al. 2015). The canonical polyadic decomposition (CPD) (see Tucker 1966; Kolda and Bader 2009) can also be used, but for computing principal components with the IVS data, the results produced by these two approaches are very similar. We use the MLSVD, which is briefly described in the Appendix, as follows.

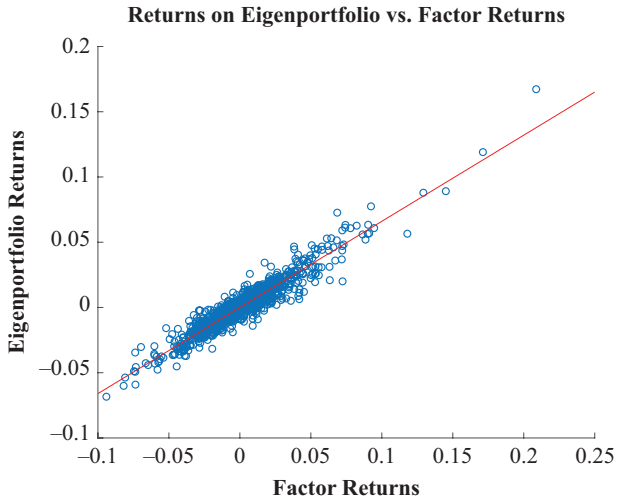
We can write the tensor R in the form

$$R = \sum_{t,i,j,k} S_{tijk}\ U_t^{(1)} \circ U_i^{(2)} \circ U_j^{(3)} \circ U_k^{(4)}, \tag{20}$$

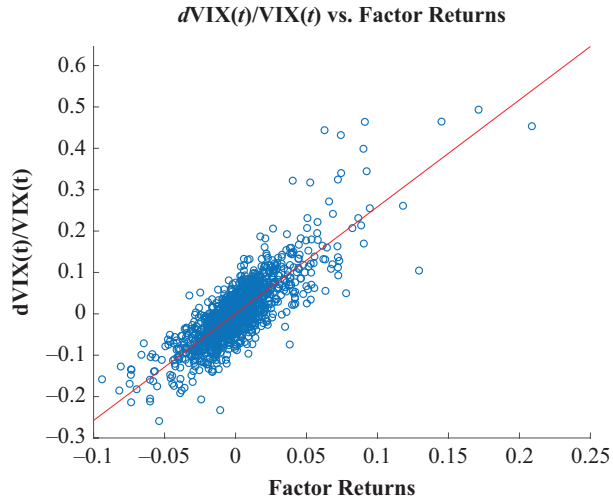
EXHIBIT 9

Output from Adapted Eigenportfolios Using a 6-Month Sliding Window

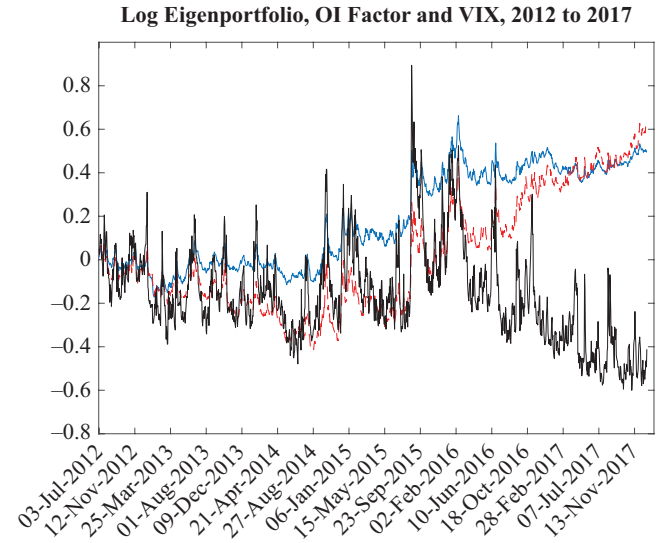
Panel A: Scatter Plot of Adapted Eigenportfolio Returns against Returns of OI-Weighted Factor Computed Using Weighting Function $\omega(\mathcal{OI}) = \mathcal{OI}$ (2012–2017)



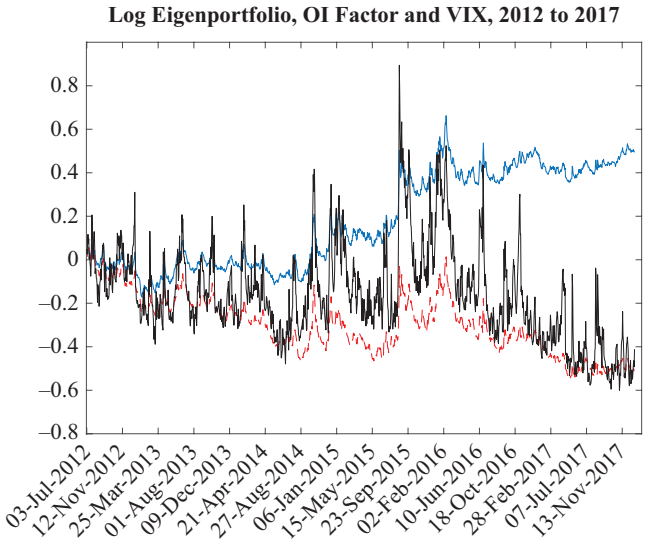
Panel B: Scatter Plot of VIX Returns against Returns of OI-Weighted Factor Computed Using Weighting Function $\omega(\mathcal{OI}) = \mathcal{OI}$ (2012–2017)



Panel C: Adapted 6-Month Sliding-Window Eigenportfolio and OI-Weighted Factor Computed Using Weighting Function $\omega(\mathcal{OI}) = \mathcal{OI}$



Panel D: Adapted 6-Month Sliding-Window Eigenportfolio and OI-Weighted Index Computed Using Weighting Function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$ (logarithm of each index)



Eigenportfolio — OI Factor — VIX

Notes: The eigenportfolio is computed using a 6-month sliding window, which is why the first 6 months of 2012 are not included in the output. By taking the OI weights to be the logarithm of 1 plus OI, we see improved tracking of the VIX compared with Panel C. Regression of the eigenportfolio onto the factor has significant intercept coefficient but improved projection error compared with the same regression with $\omega(\mathcal{OI}) = \mathcal{OI}$ (see Exhibit 10).

where $U^{(1)}$ is a $T \times T$ orthonormal matrix, $U^{(2)}$ is an $N^{(1)} \times T$ orthonormal matrix, $U^{(3)}$ is an $N^{(2)} \times N^{(2)}$ orthonormal matrix, $U^{(4)}$ is an $N^{(3)} \times N^{(3)}$ orthonormal matrix, S is a $T \times N^{(1)} \times N^{(2)} \times N^{(3)}$ real-valued tensor, and $\circ$ denotes the vector outerproduct. The main disadvantage of the MLSVD is that the so-called *core tensor* S is, in general, not diagonal or even sparse, which results in difficulties when computing a rank-d decomposition that is best in

EXHIBIT 10

Factor Loadings and Unexplained Systematic Returns for Eigenportfolio from a 6-Month Sliding Window and Various OI Weightings in Factor Construction

$\omega(\mathcal{OI})$	α_3	β_3	α_2	β_2	α_1	β_1
2012–2014						
$\mathcal{OI}$	0.0629 (1.6307)	0.6348	0.0755 (1.8777)	0.7196	0.0908 (2.2614)	0.7187
	$R^2 = 0.9387$		$R^2 = 0.9332$		$R^2 = 0.9322$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1224 (4.0669)	0.8947	0.1418 (4.4146)	1.0053	0.1745 (5.2169)	1.0000
	$R^2 = 0.9629$		$R^2 = 0.9574$		$R^2 = 0.9529$	
2013–2015						
$\mathcal{OI}$	0.0570 (1.1356)	0.5500	0.0728 (1.3413)	0.6794	0.0865 (1.5885)	0.6769
	$R^2 = 0.9303$		$R^2 = 0.9184$		$R^2 = 0.9172$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1409 (4.1438)	0.8472	0.1664 (4.4377)	0.9842	0.1934 (4.8793)	0.9745
	$R^2 = 0.9681$		$R^2 = 0.9610$		$R^2 = 0.9561$	
2014–2016						
$\mathcal{OI}$	0.0190 (0.3091)	0.5259	0.0116 (0.1746)	0.6600	0.0165 (0.2491)	0.6594
	$R^2 = 0.9183$		$R^2 = 0.9049$		$R^2 = 0.9047$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1841 (4.9527)	0.8731	0.2007 (5.1087)	0.9734	0.2184 (5.3412)	0.9687
	$R^2 = 0.9701$		$R^2 = 0.9664$		$R^2 = 0.9634$	
2015–2017						
$\mathcal{OI}$	–0.0505 (–0.8877)	0.5164	–0.0705 (–1.1757)	0.6168	–0.0803 (–1.3413)	0.6170
	$R^2 = 0.9102$		$R^2 = 0.8997$		$R^2 = 0.8992$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1705 (6.5261)	0.8926	0.1769 (6.6106)	0.9389	0.1836 (6.8505)	0.9381
	$R^2 = 0.9810$		$R^2 = 0.9800$		$R^2 = 0.9797$	
2012–2017 (all years)						
$\mathcal{OI}$	–0.0010 (–0.0289)	0.5444	–0.0020 (–0.0546)	0.6604	0.0028 (0.0781)	0.6601
	$R^2 = 0.9156$		$R^2 = 0.9033$		$R^2 = 0.9032$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1476 (7.2613)	0.8903	0.1619 (7.5857)	0.9760	0.1813 (8.2952)	0.9732
	$R^2 = 0.9697$		$R^2 = 0.9665$		$R^2 = 0.9645$	

Notes: For each sample period, there are three regressions: a three-factor regression $\frac{dE^{\mathcal{P}}}{E^{\mathcal{P}}} = \alpha_3 + \beta_3 \frac{dQ}{Q} + b_{eq} \frac{dS\&P}{S\&P} + b_{vx} \frac{dVIX}{VIX} + \varepsilon$, *a two-factor regression* $\frac{dE^{\mathcal{P}}}{E^{\mathcal{P}}} = \alpha_2 + \beta_2 \frac{dQ}{Q} + b_{eq} \frac{dS\&P}{S\&P} + \varepsilon$, *and a one-factor regression* $\frac{dE^{\mathcal{P}}}{E^{\mathcal{P}}} = \alpha_1 + \beta_1 \frac{dQ}{Q} + \varepsilon$. *The factor obtained by weighting function* $\omega(\mathcal{OI}) = \mathcal{OI}$ *is perhaps desirable because it leaves no significant excess return in the residual, but there is no reason why we should reject a factor with nonzero intercept. In contrast, the weighting function* $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$ *has significant intercept and a higher* R^2 *and hence is perhaps a better factor.*

the sense of the Frobenius norm. In practice, the rank-1 decomposition used for computing the eigenportfolio can be estimated using the top MLSVD vectors, which we write as

$$\tilde{U}^{(1)} = U_1^{(2)} \circ U_1^{(3)} \circ U_1^{(4)}, \tag{21}$$

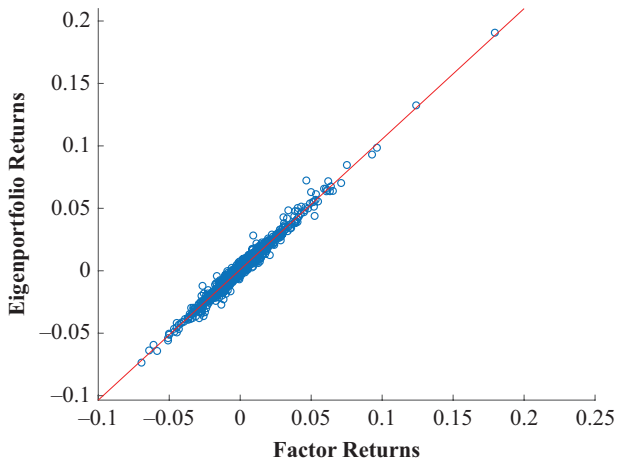
which for the IVS data is very close to what one gets using CPD. To compute the tensor eigenportfolio, we

EXHIBIT 11

Output from the Adapted Tensor Eigenportfolio Using a 6-Month Sliding Window

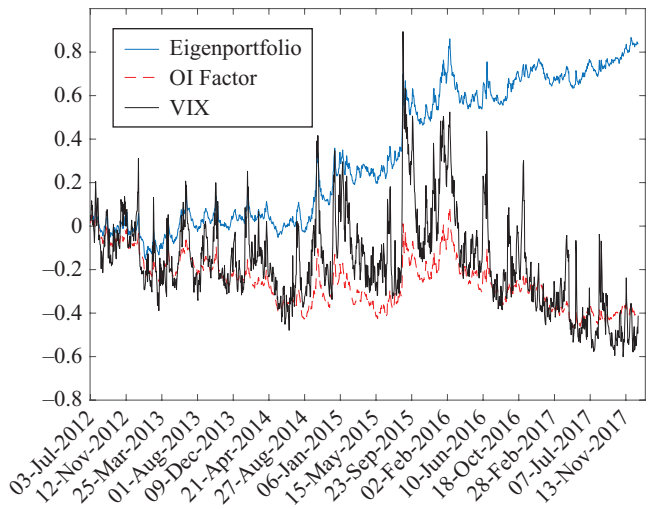
Panel A: Scatter Plot of Adapted Tensor Eigenportfolio Returns against Returns of Tensor OI-Weighted $\bar{Q}$ Factor Given by Equation 24

Returns on Eigenportfolio vs. Factor Returns



Panel B: Adapted Tensor Eigenportfolio and Tensor *OI*-Weighted Index Computed Using Weighting Function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$ (logarithm of each index)

Log Eigenportfolio, OI Factor and VIX, 2012 to 2017



Notes: Panel A is computed using weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$, from 2012 through 2017. The eigenportfolio is computed using a 6-month sliding window. The linear dependence seen here is an improvement from the linear dependence seen in Exhibit 9, Panel A, achieved using flat matrixes. Panel B does not suggest that the tensor approach to eigenportfolio and factor construction offers an improvement to simply using flat matrixes because it is roughly similar to its analogue in Exhibit 9, Panel D. Improvement as a result of tensors is more evident in Panel A and in Exhibits 12, 13, and 14.

do an elementwise division with the tensor of standard deviations:

$$\pi^{(1)}_{ijk} \propto \frac{\tilde{U}^{(1)}_{ijk}}{h_{ijk}}. \tag{22}$$

It remains to be decided how to normalize this $\pi^{(1)}$. If we normalize with one global summation over the multi-index (i, j, k), there would have been no need to use tensors. However, if there is a dimension such that each of the index values has a different corresponding OI-weighted factor, then the normalization needs to be done separately for the different values of this index. The logic for using different normalizations for different index values will become clearer in the following example.

Example: Different Factors for Each Maturity

Suppose that we construct a different OI-weighted factor for each maturity, for a total of eight factors. Moreover, suppose that our factor considers the β for each option to be its loading on the factor corresponding to the same maturity; that is, there are factors Q_j for $j = 1, 2, \ldots, 8$, with $dQ_j(t)/Q_j(t) \propto \sum_{i,k} \omega(\mathcal{OI}_{ijk}(t)) r_{ijk}(t)$, where $\mathcal{OI}$ is the tensor of open interests. The factor model for returns is then

$$r_{ijk}(t) = \beta_{ijk} \frac{dQ_j(t)}{Q_j(t)} + \xi_{ijk}(t),$$

where ξ is a tensor of idiosyncratic noise that is independent of the factor processes in Q_j. In this case, for the β_{ijk}/h^2_{ijk}'s to line up with the eigenportfolio in Equation 22, the normalization needs to be done as follows for each j:

$$\begin{aligned} \pi^{(1)}_{ijk} &= \frac{\tilde{U}^{(1)}_{ijk}/h_{ijk}}{\sum_{i,k} \tilde{U}^{(1)}_{ijk}/h_{ijk}} \\ \frac{dQ_j(t)}{Q_j(t)} &= \frac{\sum_{i,k} \omega(\mathcal{OI}_{ijk}(t)) r_{ijk}(t)}{\sum_{i,k} \omega(\mathcal{OI}_{ijk}(t))}. \end{aligned} \tag{23}$$

EXHIBIT 12

Factor Loadings and Systematic Unexplained Returns for Tensor Eigenportfolio from 6-Month Sliding Window and Various OI Weightings in Tensor-Factor Construction

$\omega(\mathcal{OI})$	α_3	β_3	α_2	β_2	α_1	β_1
2012–2017 (all years)						
$\mathcal{OI}$	0.0891 (3.0439)	0.7384	0.0923 (2.9875)	0.8432	0.0925 (3.0107)	0.8432
	$R^2 = 0.9466$		$R^2 = 0.9404$		$R^2 = 0.9404$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$	0.2092 (11.8532)	0.9673	0.2211 (11.8433)	1.0466	0.2339 (12.3862)	1.0450
	$R^2 = 0.9806$		$R^2 = 0.9782$		$R^2 = 0.9775$	

Notes: For each sample period, there are three regressions: a three-factor regression $\frac{dE^{\varphi}}{E^{\varphi}} = \alpha_3 + \beta_3 \frac{d\bar{Q}}{\bar{Q}} + b_{eq} \frac{dS\&P}{S\&P} + b_{vx} \frac{dVIX}{VIX} + \varepsilon$, *a two-factor regression* $\frac{dE^{\varphi}}{E^{\varphi}} = \alpha_2 + \beta_2 \frac{d\bar{Q}}{\bar{Q}} + b_{eq} \frac{dS\&P}{S\&P} + \varepsilon$, *and a one-factor regression* $\frac{dE^{\varphi}}{E^{\varphi}} = \alpha_1 + \beta_1 \frac{d\bar{Q}}{\bar{Q}} + \varepsilon$. *The tensor regressions shown in this exhibit have higher* R^2 *than their counterparts in Exhibit 10, which is an indication that the tensor approach to factor construction is better.*

Finally, the global factor (to compare with the eigenportfolio) is simply the mean of the tenor factors:

$$\frac{d\bar{Q}(t)}{\bar{Q}(t)} = \frac{1}{N^{(2)}} \sum_j \frac{dQ_j(t)}{Q_j(t)}. \tag{24}$$

Exhibit 11 shows improved results if this maturity-wise approach is used with the normalizations in Equation 23 along with the weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$. The linear dependence between the tensor eigenportfolio and the $\bar{Q}$ seen in Exhibit 11, Panel A, is a considerable improvement from the linear dependence shown in Exhibit 9, Panel A, which was obtained using flat matrixes. The ex post regression diagnostics of the tensor approach show an improvement over the flat matrix approach; Exhibit 12 lists regression outputs that have higher a R^2 (i.e., less projection error) than their counterparts in Exhibit 10.

Finally, we show the in-sample plots of cumulative returns for the tensor portfolios next to the cumulative returns of the portfolios from the flat matrixes; all portfolios are constructed with weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$. Exhibits 13 and 14 show these results, from which it is clear that the tensor factor constructed in this subsection allows for eigenportfolio tracking that is much closer to the factor. This is evidence that the family of OI-based tensor factors comprises a good factor, in the sense that it can track the eigenportfolio returns. We recall that we are concerned with the eigenportfolio when constructing factors because we know from theoretical analysis of the spike model that the first eigenportfolio will have weights close to the vector $h^{-2}\beta$, where β are loadings on a dominant factor. Hence, to determine whether we have a dominant factor, we should make comparisons with the eigenportfolio, and in the tensor IVS data context the results come out better, most likely because the data are heterogeneous. It is therefore beneficial to respect the tensor structure, which MLSVD does.

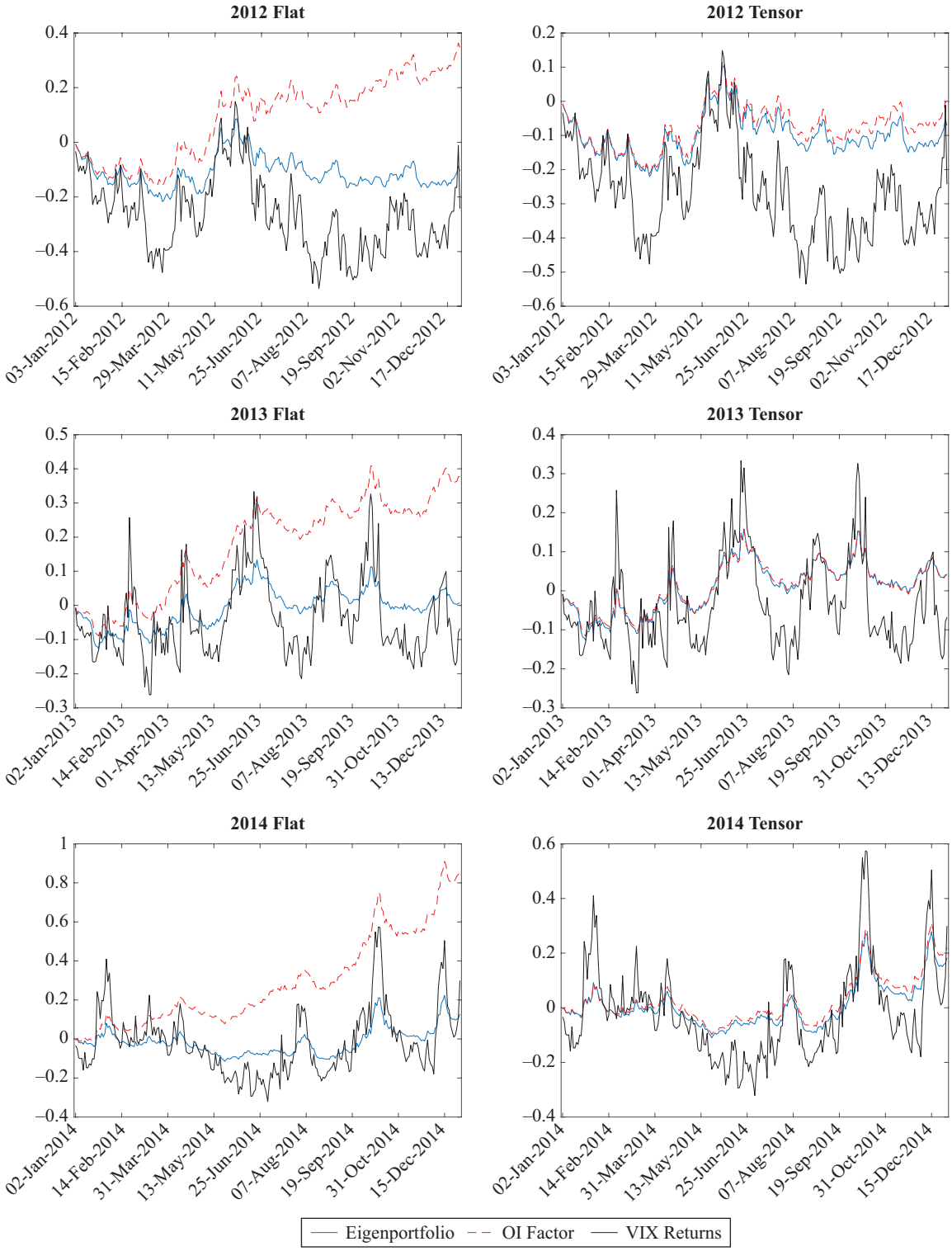
SUMMARY AND CONCLUSION

We have carried out PCA of a dataset of IVSs from options on US equities. The analysis contains daily implied volatilities for the 6 years 2012–2017 for 500 equities (names), eight maturities, and seven strikes (i.e., 28,000 data points daily for 251 × 6 days). We have posed and answered three questions in this data-driven study: (1) What is the essential information in this dataset, or how many factors are needed to represent the data with the residual being noise; (2) by analogy with equity-returns PCA analysis, can the principal eigenportfolio be associated with a market portfolio, and what should that market portfolio be; and (3) because the natural structure of the data is that of a four-dimensional tensor (time, name, maturity, strike), do we benefit from preserving this structure in the PCA analysis—that is, by doing a tensor PCA?

The analysis of question 1 is in the section Matrix of Implied Volatility Returns. Building on the work of Avellaneda and Dobi (2014), we perform matrix PCA on the flattened IVS data excluding stock returns and arrive at the conclusion that the number of significant factors is nine. For comparison, the number of significant

EXHIBIT 13

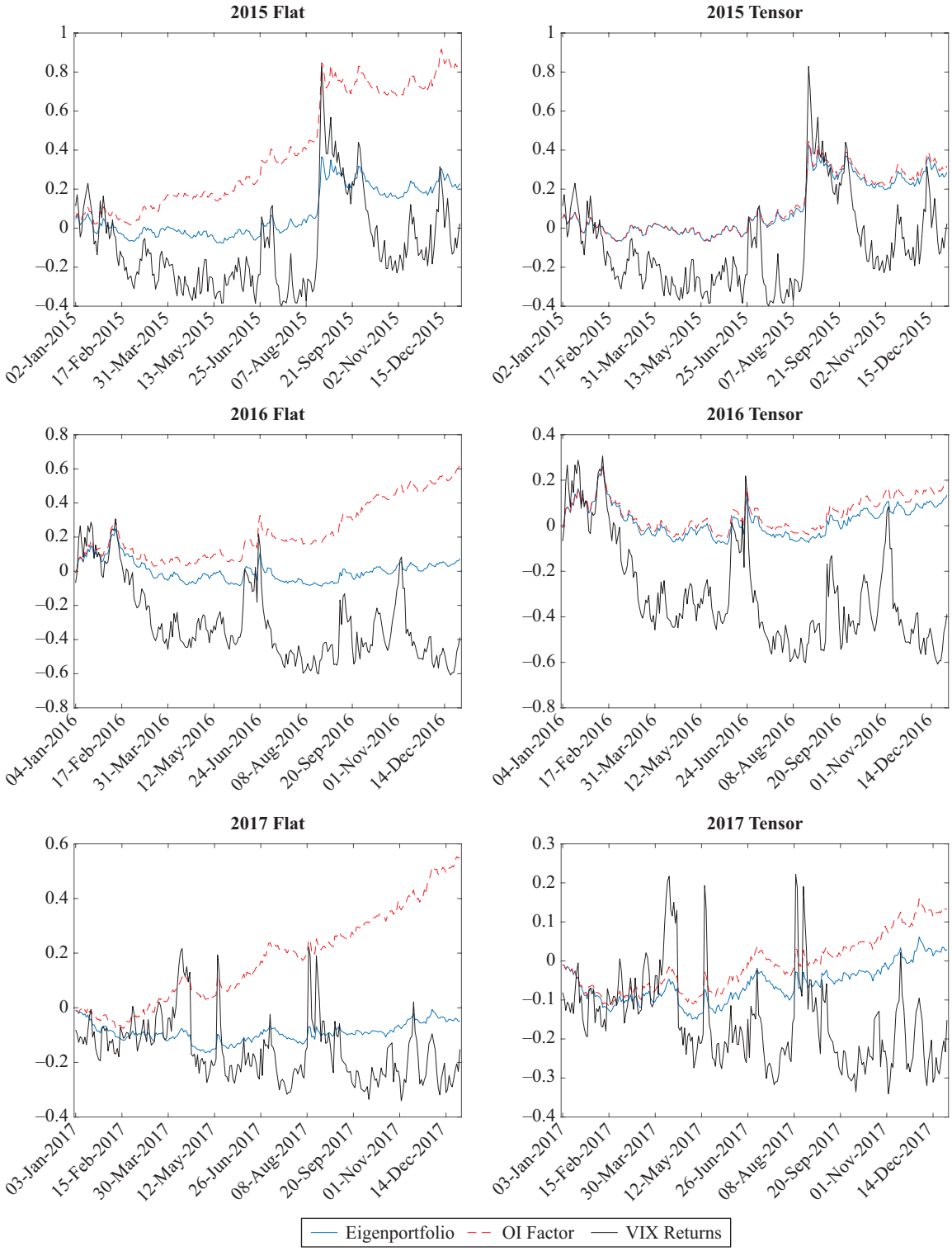
Comparison of the Eigenportfolio Tracking the OI Factor for Both Flat Matrixes (on the left) and Tensors (on the right)



Notes: All portfolios are constructed with weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$. *The plots display the logarithm of each index. These are in-sample fits for each of the years 2012–2014.*

EXHIBIT 14

Comparison of Eigenportfolio Tracking OI Factor for Both Flat Matrixes (on the left) and Tensors (on the right)



Notes: All portfolios are constructed with weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$. *The plots display the logarithm of each index. These are in-sample fits for each of the years 2015–2017.*

factors for equity returns is normally in the range of 15 to 20. There is a very large dimension reduction to the IVS dataset because of the high degree of structure in option prices and hence in the IVS. In this section, we also introduced the concept of effective dimension for the residual because, contrary to theoretical factor analysis and equity returns analysis, the IVS residuals retain structural patterns or correlations. We find that the effective spatial dimension (name, maturity, strike) of the residuals is closer to 500 than to the nominal dimension of 28,000. This is the way spectral analysis (i.e., PCA) deals with the heterogeneity of the IVS data: The residuals have patterns even after the market information has been taken out.

The analysis of question 2 is in the section Principal Eigenportfolios and OI-Weighted Indexes. The main result here is the construction of the analog for IVS of the market portfolio for equity returns. This is based on using the open interest of the options and their (unitless) Vega. OI is the number of contracts on a given day for each name, maturity, and strike, and the Vega is an associated sensitivity to volatility. We find that portfolios of IVS returns weighed suitably by OI and Vega do track the IVS eigenportfolio. This provides an interpretation of the eigenportfolio that is analogous to the one for equity returns, is robust, and can be used in ways the VIX, the SPX volatility index, is used.

The analysis of question 3 can be found in the Factors and Eigenportfolios Using Tensors section. Yes, we find that retaining the tensor structure in the eigenportfolio analysis makes a difference in that the OI–Vega-weighted (tensor) IVS returns portfolio tracks the (tensor) eigenportfolio much better. This is a strong indication that data structure matters and data flattening should be avoided if possible.

There are obviously many, many more questions that can and should be asked about the IVS dataset, including theoretical ones about the methodology used. It is an evolving research enterprise.

Appendix A

DESCRIPTION OF DATA

Implied Volatility

Implied volatilities are not directly observable in the market and must be derived from the prices of traded options. At any point in time, for a given underlying stock, a variety of call and put options will be available to trade on the market, each of which will have a specific strike price and time to maturity. Given the observed prices of these options, it is then possible to infer an implied volatility using a numerical method.

Options on individual stocks have an American-style exercise feature and must be priced using a numerical algorithm because no closed-form solution is available. For this purpose, OptionMetrics uses the industry standard Cox–Ross–Rubinstein binomial tree model. This model can accommodate underlying securities with either discrete dividend payments or a continuous dividend yield. An option is priced by working backward through the tree from the maturity date when the payoff is known and incorporating any potential value arising from the possibility of early exercise at each node. The calculated price of the option at time $t = 0$ is the model price.

To compute the implied volatility of an option given its price, the model is run iteratively with different values of σ until the model price of the option converges to its market price, defined as the midpoint of the option's best closing bid and best closing offer prices. At this point, the final value of σ is the option's implied volatility.

This model can be adapted to account for the discrete dividends that stocks typically pay on a quarterly basis. The approach taken is to adjust the price of the underlying stock by subtracting the discounted value of all dividends to be paid between now and the expiry of the option.

Once the implied volatility has been calculated for an option, it is a simple matter to calculate its delta using the Black–Scholes model. We can then create a grid of time to maturity versus delta with a corresponding implied volatility where it is known. This will result in a grid of implied volatilities that have very different times to maturity and deltas from those needed to form a standardized grid, which is referred to as an IVS.

OptionMetrics calculates its standardized option-implied volatilities using a kernel-smoothing technique. The data are first organized by the log of days to expiration and by call-equivalent delta (delta for a call and one plus delta for a put). A kernel smoother is then used to generate a smoothed volatility value at each of the specified interpolation grid points. At each grid point j on the volatility surface, the smoothed volatility $\hat{\sigma}_j$ is calculated as a weighted sum of option-implied volatilities:

$$\hat{\sigma}_j = \frac{\sum_i \mathcal{V}_i \sigma_i \Phi(x_{i,j}, y_{i,j}, z_{i,j})}{\sum_i \mathcal{V}_i \Phi(x_{i,j}, y_{i,j}, z_{i,j})},$$

where i is indexed over all the options for that day, $\mathcal{V}_i$ is the Vega of the option, σ_i is the implied volatility, and Φ is the kernel function:

$$\Phi(x, y, z) = \frac{1}{\sqrt{2\pi}} e^{-\left(\left(\frac{x^2}{2h_1}\right)+\left(\frac{y^2}{2h_2}\right)+\left(\frac{z^2}{2h_3}\right)\right)}$$

EXHIBIT A1
Bucketing Scheme for Open Interest

τ Bucket (days)	Min Value (days)	Max Value (days)
30	1	45
60	46	75
91	76	106
122	107	137
152	138	167
182	168	227
273	228	319
365	320	

where

$$x_{i,j} = \log(T_i / T_j)$$
$$y_{i,j} = \Delta_i - \Delta_j$$
$$z_{i,j} = I_{CP_i = CP_j}.$$

Values $x_{i,j}$, $y_{i,j}$ and $z_{i,j}$ are measures of the distance between the option and the target grid point, $T_i(T_j)$ is the number of days to the expiration of the option (grid point), $\Delta_i(\Delta_j)$ is the call-equivalent delta of the option (grid point), $CP_i(CP_j)$ is the call/put identifier of the option (grid point), and I_0 is an indicator function. The kernel bandwidth parameters were chosen empirically and are set as $h_1 = 0.05$, $h_2 = 0.005$, and $h_3 = 0.001$.

Open Interest

Unlike implied volatility, open interest is directly observable in the market and represents the number of contracts in a particular option (underlying stock, strike, and expiry date) open at a point in time. For each open contract, there will be a party who is long the option and conversely one who is short the option. Open interest is therefore concentrated in the most popular contracts, which tend to be those closest to at-the-money and have less than one year but more than two weeks to expiry.

OptionMetrics provides an open interest for every available contract at the close of each trading day because this is the information available from the exchange. However, as described earlier, we are interested in the open interest at the grid points we have chosen for our IVSs, namely constant Δs and constant maturities. To estimate the open interest for these points, we used a bucketing approach.

First, we allocated each option to one of eight time buckets according to its days to expiry (τ), where the limits on these ranges are as given in Exhibit A1. Similarly, we allocated each option to one of eight Δ buckets according to where the limits on these ranges are as given in Exhibit A2. Once this was done, each option was allocated to one of our 56 grid points, and we then summed the open interest across all options at each grid point for each stock on each day in our dataset.

EXHIBIT A2
Bucketing Scheme for Delta

Δ Bucket	Min Value	Max Value
–20	–25	–1
–30	–35	–24
–40	–45	–34
50	45	54
40	35	44
30	25	34
20	1	24

APPENDIX B

THE MULTILINEAR SVD

If A is a matrix,[6] we can use the singular value decomposition (SVD) to expand

$$A = A(i,t), \;\; i = 1, \ldots, N, \;\; t = 1, \ldots, T$$

in the form

$$A = U \Sigma V^*.$$

Here U is an $N \times N$ unitary matrix, V is a unitary $T \times T$ matrix, and Σ is a diagonal matrix with entries $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_R \geq 0$, where the index R is the rank of A and $R \leq \min\{N, T\}$. In terms of the components, we have

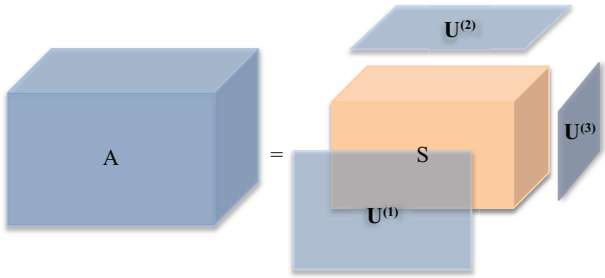
$$A(i,t) = \sum_{r=1}^{R} \sigma_r U(i,r) V(t,r), \; i = 1, \ldots, N, \; j = 1, \ldots, T.$$

For this expansion, only R columns of the unitary matrixes U and V are needed to represent the matrix A. When the sum stops at $\bar{R} < R$, the corresponding matrix $A_{\bar{R}}$ is the best rank $\bar{R}$ approximation of the matrix A in the Frobenius norm.

These properties do not generalize to tensors except in special and rather limited ways. A commonly used approach

[6] The notation in this Appendix is a bit different from that used earlier in the article and closer to that used in the linear algebra literature.

Exhibit B1
Core Decomposition of a Three-Dimensional Cube



for tensors is the MLSVD, or Tucker decomposition, which for a fourth-order tensor has the form:

$$A(\underline{i},t) = \sum_{\underline{1} \le \underline{i}' \le \underline{N}, 1 \le t' \le T} S(\underline{i}',t')U(\underline{i},\underline{i}')V(t,t'). \tag{B1}$$

Here, $\underline{i} = (i_1, i_2, i_3)$, $\underline{N} = (N_1, N_2, N_3)$, V is a unitary $T \times T$ matrix, and U is a unitary tensor of the form

$$U(\underline{i},\underline{i}') = U^{(1)}(i_1,i_1')U^{(2)}(i_2,i_2')U^{(3)}(i_3,i_3') \tag{B2}$$

and the matrixes $U^{(1)}$, $U^{(2)}$, $U^{(3)}$ are unitary of size $N_1 \times N_1$, $N_2 \times N_2$, and $N_3 \times N_3$, respectively. The fourth-order tensor S is not diagonal anymore, in general. However, it has the property of *all orthogonality*:

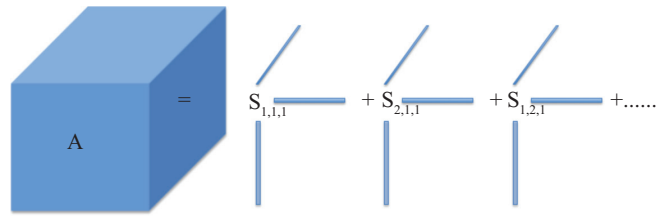
$$\sum S(\underline{i},t)S(\underline{i}',t') = 0 \tag{B3}$$

where the indexes $(\underline{i},t)$ and $(\underline{i}',t')$ are equal except for one of the four components and the sum is over the three equal components. This means that distinct slices of each orientation (Horizontal [N_1], Vertical [N_2], Frontal [N_3], Time [T]) are orthogonal. Moreover, the indexes can be permuted so that the sums of squares over all except for one index are ordered by size

$$\begin{aligned}
\sigma_1^H &= \sqrt{\sum_{j,k,t} S^2(1,j,k,t)} \ge \sigma_2^H \ge \cdots \ge \sigma_{N_1}^H \\
\sigma_1^V &= \sqrt{\sum_{i,k,t} S^2(i,1,k,t)} \ge \sigma_2^V \ge \cdots \ge \sigma_{N_2}^V \\
\sigma_1^F &= \sqrt{\sum_{i,j,t} S^2(i,j,1,t)} \ge \sigma_2^F \ge \cdots \ge \sigma_{N_3}^F \\
\sigma_1^T &= \sqrt{\sum_{i,j,k} S^2(i,j,k,1)} \ge \sigma_2^T \ge \cdots \ge \sigma_T^T
\end{aligned} \tag{B4}$$

where we write here $\underline{i} = (i, j, k)$.

Exhibit B2
Tucker Decomposition



The MLSVD representation (Equation B1) is exact and is obtained from the application of SVDs to all the possible flattenings of the tensor and organizing the output suitably to get the result (see De Lathauwer, De Moor, and Vandewalle 2000; Kolda and Bader 2009; and Cichocki et al. 2015). A schematic of the MLSVD expansion is shown in Exhibit B1. Another expansion, not used here, is the canonical polyadic decomposition that is based on Frobenius norm minimization, a nonconvex problem; a schematic is shown in Exhibit B2.

In tensor PCA, we are interested in the covariance over time of the data, which we assume here is already normalized. In the notation of this Appendix, we define this covariance as

$$C(\underline{i},\underline{j}) = \sum_t A(\underline{i},t)\overline{A(\underline{j},t)}. \tag{B5}$$

Using the MLSVD representation of A in Equation B1, we deduce the representation

$$C(\underline{i},\underline{j}) = \sum_{\underline{i}',\underline{j}'} s(\underline{i}',\underline{j}')U(\underline{i},\underline{i}')\overline{U(\underline{j},\underline{j}')} \tag{B6}$$

where

$$s(\underline{i},\underline{j}) = \sum_t S(\underline{i},t)S(\underline{j},t). \tag{B7}$$

Using the unitarity of the tensor U (Equation B2) in the covariance expansion (Equation B6) we obtain

$$\sum_{\underline{j}} C(\underline{i},\underline{j})U(\underline{j},\underline{k}) = \sum_{\underline{i}'} s(\underline{i}',\underline{k})U(\underline{i},\underline{i}'). \tag{B8}$$

We see from this expression that the unitary tensor U is not an eigenvector of the covariance tensor C in Equation B8 because s is not diagonal in general. As noted earlier, U has the form

$$U(\underline{i},\underline{j}) = U^{(1)}(i_1,j_1)U^{(2)}(i_2,j_2)U^{(3)}(i_3,j_3)$$

with $U^{(1)}(i_1, j_1)$, $U^{(2)}(i_2, j_2)$, $U^{(3)}(i_3, j_3)$ unitary matrixes of size $N_1 \times N_1$, $N_2 \times N_2$, $N_3 \times N_3$, respectively. Equation B8 captures rather clearly the scope the MLSVD representation of A in Equation B1 in that, although it is not a spectral form for the covariance, the tensor s on the right has positive diagonal elements that can be ordered and has off-diagonal elements that are often small because of the total orthogonality property (Equation B3); however, this has to be verified separately and it is not generally true. When, however, $s(\underline{1}, \underline{1})$ is large, then we can take

$$\tilde{U}^{(1)}(\underline{i}, \underline{1}) = U^{(1)}(i_1, 1),\ U^{(2)}(i_2, 1),\ U^{(3)}(i_3, 1) \tag{B9}$$

as the principal tensor eigenvector, as we did with a slightly different notation in Equation 21.

We close this Appendix with a few comments that complement the discussion up to now.

First, the truncation that produces the principal tensor eigenvector does not, in general, arise from a Frobenius norm minimization of the difference $\|A - A_{PCA}\|$. However, for the IVS data, the principal tensor eigenvector using MLSVD as described here and using CPD by Frobenius norm minimization produces essentially the same result. There will be a difference for multifactor tensor PCA, an issue that is not considered here.

Second, it is observed in practice that if there is a big gap in the size of the sigmas (see Equation B4), then this kind of truncation does behave like it does for matrixes. That is, the residual tends to behave as if it came from random entries.

Third, how does one test that a tensor has entries that behave as if they are random? We saw in the matrix case in the Matrix of Implied Volatility Returns section that one looks at the histogram of the singular values and compares this with a suitably adapted MP density—using a KS test, for example. For tensors, a test can be developed by using matrix flattenings of the tensor (e.g., horizontal, vertical, frontal). Because the MP law is an asymptotic one, care must be taken that the data structure is constructed so that it gives rise to flattenings for which this asymptotic law can actually be used. This is work in progress at present.

Fourth, there are other methods for constructing tensor principal components. Alternating least squares (or alternating SVDs) are often used, but little can be said theoretically, and direct (nonconvex in general) Frobenius norm optimizations using gradient descent or stochastic gradient descent avoid getting stuck in local minima early on.

REFERENCES

Avellaneda, M. "Hierarchical PCA and Applications to Portfolio Management." Working paper, New York University Courant, 2019.

Avellaneda, M., D. Boyer-Olson, J. Busca, and P. Friz. 2002. "Reconstruction of Volatility: Pricing Index Options by the Steepest Descent Approximation." *Risk*, October: 91–95.

Avellaneda, M., and D. Dobi. "Modeling Volatility Risk in Equity Options Market: A Statistical Approach." Working paper, New York University Courant, 2014.

Avellaneda, M., and J. H. Lee. 2010. "Statistical Arbitrage in the US Equities Market." *Quantitative Finance* 10 (7): 761–782.

Benaych-Georges, F., and R. R. Nadakuditi. 2011. "The Eigenvalues and Eigenvectors of Finite, Low Rank Perturbations of Large Random Matrices." *Advances in Mathematics* 227 (1): 494–521.

Berman, M. 2019. "Improved Estimation of the Intrinsic Dimension of a Hyperspectral Image Using Random Matrix Theory." *Remote Sensing* 11 (9): 1049.

Boyle, P. 2014. "Positive Weights on the Efficient Frontier." *North American Actuarial Journal* 18 (4): 462–477.

Cichocki, A., D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. 2015. "Tensor Decompositions for Signal Processing Applications: From Two-Way to Multiway Component Analysis." *IEEE Signal Processing Magazine* 32 (2): 145–163.

Cont, R., and J. Da Fonseca. 2002. "Dynamics of Implied Volatility Surfaces." *Quantitative Finance* 2 (1): 45–60.

De Lathauwer, B., L. De Moor, and J. Vandewalle. 2000. "A Multilinear Singular Value Decomposition." *SIAM Journal on Matrix Analysis and Applications* 21 (4): 1253–1278.

Dobi, D. "Modeling Volatility Risk in Equity Options: A Cross-Sectional Approach." Ph.D. dissertation, New York University, 2014.

El Karoui, N. 2008. "Spectrum Estimation for Large Dimensional Covariance Matrices Using Random Matrix Theory." *Annals of Statisics* 36 (6): 2757–2790.

Fama, E. F., and K. R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance*, 47 (2): 427–465.

Griffin, J. M., and A. Shams. 2017. "Manipulation in the VIX?" *The Review of Financial Studies* 31 (4): 1377–1417.

Johnstone, I. M., and D. Paul. 2018. "PCA in High Dimensions: An Orientation." *Proceedings of the IEEE* 106 (8): 1277–1292.

Kolda, T. G., and B. W. Bader. 2009. "Tensor Decompositions and Applications." *SIAM Review* 51 (3): 455–500.

Ledoit, O., and S. Péché. 2011. "Eigenvectors of Some Large Sample Covariance Matrix Ensembles." *Probability Theory and Related Fields* 151 (1–2): 233–264.

Ledoit, O., and M. Wolf. 2004. "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis* 88 (2): 365–411.

——. 2012. "Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices." *The Annals of Statistics* 40 (2): 1024–1060.

Markowitz, H. 1952. "Portfolio Selection." *The Journal of Finance* 7 (1): 77–91.

Plerou, V., P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley. 2002. "Random Matrix Approach to Cross Correlations in Financial Data." *Physical Review E* 65 (6): 066126.

Roll, R., and S. A. Ross. 1980. "An Empirical Investigation of the Arbitrage Pricing Theory." *The Journal of Finance* 35 (5): 1073–1103.

Tucker, L. R. 1966. "Some Mathematical Notes on Three-Mode Factor Analysis." *Psychometrika* 31 (3): 279–311.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Neural Networks in Finance: *Design and Performance*

Irene Aldridge and Marco Avellaneda

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/4/39

ABSTRACT: *Neural networks have piqued the interest of many financial modelers, but the concrete applications and implementation have remained elusive. This article discusses a step-by-step technique for building a potentially profitable financial neural network. The authors also demonstrate a successful application of the neural network to investing based on daily and monthly financial data. The article discusses various components of neural networks and compares popular neural network activation functions and their applicability to financial time series. Specifically, use of the tanh activation function is shown to closely mimic financial returns and produce the best results. Incorporating additional inputs, such as the S&P 500 prices, also helps improve neural networks' forecasting performance. Longer training periods deliver strategies that closely mimic common technical analysis strategies, such as moving-average crossovers, whereas shorter training periods deliver significant forecasting power. The resulting neural network–based daily trading strategies on major US stocks significantly and consistently outperform the buy-and-hold positions in the same stocks.*

Evaluating Cost Efficiencies in Asset Management and Custody Banks Using Data Envelopment Analysis

D. K. Malhotra, Rashmi Malhotra, and Robert Nydick

D. K. Malhotra
is a professor of finance at Thomas Jefferson University in Philadelphia, PA.
davinder.malhotra@jefferson.edu

Rashmi Malhotra
is a professor of decision and systems sciences at Saint Joseph's University in Philadelphia, PA.
rmalhotr@sju.edu

Robert Nydick
is a professor of management and operations at Villanova University in Villanova, PA.
robert.nydick@villanova.edu

KEY FINDINGS

- Initially most asset management firms exhibit increasing returns to scale (IRS) until they become too large.
- As companies expand in terms of total assets and total revenues, they do reach a point at which size is too large and cost inefficiencies manifest in the form of decreasing returns to scale.
- Whenever asset management and custody banks reach their most productive scale size and begin to experience decreasing returns to scale, they make structural changes by adopting new technologies and/or changes in the processes so that their total operating expenses decline and they can grow again without a proportionate increase in operating costs.

ABSTRACT: *As the aging baby boomer generation saves for retirement and demands suitable retirement savings vehicles, the role of asset and custody management banks is growing and the assets under their management are rising. The authors used a data envelopment analysis (DEA) model to evaluate the economies of scale in the asset management and custody bank subindustry of the capital market industry for the period of 2010 to 2017. They measured company size in two ways: total assets under management of the firm and total revenue generated by the firm. The authors evaluated cost efficiencies with cost measured in terms of total expenses, and they evaluated two important components of total expenses: total distribution expenses and general administrative and selling expenses. Using a returns to scale (RTS) DEA model, they found that most asset management firms initially exhibit increasing RTS until they become too large. The trend persisted when general and administrative expenses were evaluated and when total distribution expenses were evaluated.*

TOPICS: ***Portfolio management/multi-asset allocation, portfolio construction, wealth management, retirement****

The asset management and custody bank subindustry (which is part of the capital market industry) includes mutual funds, closed-end funds, unit investment trusts, investment counseling firms, and other groups that manage the pooled savings of individuals or organizations. Asset management and custody banks are one of the two primary subindustries within the capital market industry and account for more than 57% of the industry in terms of market capitalization.[1] According to the Investment Company Institute, the asset management

[1] S&P Capital IQ Industry Surveys: Capital Markets, 2018.

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

industry will likely continue to see substantial growth over the next couple of decades as the aging baby boomer generation (approximately 77 million Americans born between 1946 and 1964) saves for retirement and demands more suitable retirement savings vehicles.

Asset management companies compete for the business of potential investors by offering various types of investment products that match the investor's risk–return profile at a reduced cost. Because larger size enables firms to reduce costs to potential investors, size has become a source of increasing competitive advantage for the asset management industry. At the same time, as global capital markets have expanded again in the aftermath of the economic crisis of 2008–2009, regulatory oversight that increases capital requirements for liquidity to prevent a repeat of the crisis has also given larger firms an advantage. Compared to smaller firms, larger firms are better positioned and equipped to take advantage of the growing market by leveraging their global scale and market share, in terms of higher revenue, because assets are the resources that firms use to generate revenue. Meanwhile, the asset management industry is going through structural changes with the growing importance of algorithmic trading, artificial intelligence–based portfolio trading, technological innovations, and back office automation that is a causing a decline in fees and expense ratios.

Investment companies managed 9,356 mutual funds in the United States as of December 2017, with $18.3 trillion in assets under management, up from $11.8 trillion at the end of 2010, according to the Investment Company Institute. Although the dollar amount of assets managed by the asset management industry has increased several-fold, to our knowledge, no empirical research exists on economies of scale in the asset management and custody bank subindustry. This issue of economies of scale in asset management and custody banks is important for three reasons. First, previous studies have argued that expenses have a direct impact on returns to investors. If economies of scale exist, then with every increase in size, expenses will decrease and return to investors will increase. This issue is even more important in the new regulatory environment, in which risk taking is regulated. Second, recent studies in the area of mutual funds have shown that individual investors pay attention to fund expenses, and net fund flows are influenced by fund costs. Therefore, the asset management and custody bank industry should understand how expenses and size affect returns to investors. Third, as slim operating margins come to characterize the custody banking business, the industry is extremely concentrated and participants are always looking to improve profitability through acquisitions to achieve greater economies of scale. The four largest custody banks, all based in the United States, are responsible for nearly half of total global custody assets.

Numerous factors have contributed to significant growth and competition in the asset management and custodian bank industry during the time after the 2008–2009 economic crisis, according to the Office of Comptroller of Currency. First, there has been tremendous growth in the assets under management in the financial services industry. Second, the globalization of capital markets has contributed to the smooth flow of funds across international borders, and this has opened up a new pool of investors and new opportunities for investments. Third, mutual funds, pension funds, and similar financial instruments have led to a proliferation of investment alternatives. Furthermore, strong economic performance has increased assets under management in the industry, and assets are only expected to increase further.

To better understand the role of assets in the future, the asset management industry and regulatory agencies need to gain a better understanding of how the size of investment and asset management firms affects the costs incurred. If an asset management company experiences increasing returns to scale (IRS), the company can still grow and experience a less-than-proportionate increase in expenses and enhance the rate of return for its clients. On the other hand, if a company experiences decreasing returns to scale (DRS), the company has attained its optimal size and should not expand further. If the company wants to expand further, it should refine its business processes so that its efficiency improves and it can experience IRS. The contribution of this article is to develop a more efficient model for the asset management and custody bank industry to promote smooth exchange of capital among investors. We use an RTS data envelopment analysis (DEA) model to benchmark the different asset management and custody banks in the industry. In addition, we evaluate the most productive scale size for asset management and custody banks.

To gain a better understanding of how mergers and acquisitions in the asset management and custodian bank industry affect expenses and benefit fund shareholders

in the form of higher returns and increased shareholder wealth, we extended previous studies by evaluating the scale efficiency of US asset management and custody banks. Our study examines the economies of scale of the US asset management and custody banks over a period of eight years of significant growth (2010 to 2017), long enough to investigate the consistency of economies of scale. To our knowledge, no study has so extensively examined scale efficiency for this subindustry of the capital market industry.

LITERATURE REVIEW

Recently, many studies have illustrated how the DEA model, a nonparametric methodology, can be used to analyze different aspects of business entities. DEA has traditionally been used to assess the comparative efficiency of homogenous operating units, such as schools, hospitals, utility companies, sales outlets, prisons, and military operations.

Studies Focused on Benchmarking Banks

Malhotra, Poteau, and Malhotra (2012) developed a multidimensional framework using DEA to benchmark the performance of 35 commercial banks in India. The researchers compared the relative performance of the banks against one another using eight performance variables (return on assets, interest income relative to total funds, interest spread, interest expense to interest earned ratio, asset utilization ratio, efficiency ratio, debt leverage, and capital adequacy ratio as the benchmark parameters) from 2005 to 2009. They also looked at the variations in the performance of these banks over that time period to assess their progress toward efficiency.

Kao and Liu (2004) computed efficiency scores based on the data contained in the financial statements of 24 commercial banks in Taiwan to make advanced predictions about their performance. They reported that all of the efficiency scores calculated from data in subsequent financial statements were within the efficiency score ranges their model forecast, including even the bad performance of the two banks taken over by the Financial Restructuring Fund of Taiwan.

Pille and Paradi (2002) developed models to detect weaknesses in credit unions in Ontario, Canada. They compared four DEA models with the equity to asset ratio and with the government regulator's extensively modified "Z-score" model and found that the equity/asset ratio predicted failure as well as any of the other models; the ratio was not improved upon by the much more complex Z-score model.

Halkos and Salamouis (2004) explored the efficiency of Greek banks using a number of suggested financial efficiency ratios for the time period of 1997–1999. They showed that DEA could be used as either an alternative or a complement to ratio analysis to evaluate an organization's performance. The efficiency of the DEA model increased as the size of the total assets increased.

Neal (2004) used DEA and Malmquist productivity indexes to investigate X-efficiency and productivity change in Australian banking between 1995 and 1999. The study differed from earlier studies in that it examined efficiency by bank type. The researchers found that regional banks were less efficient than other bank types. They concluded that diseconomies of scale occurred very early and hence were not a sufficient basis on which to allow mergers between large banks to proceed.

Studies Focused on Analyzing the Branch Efficiencies

Balfour et al. (2015) measured the performance and relative operating efficiency of mini branches of a commercial bank. They concluded that the number of full-time employees and the amount of service charges were statistically significant in explaining the efficiency of the bank branches. Paradi and Schaffnit (2004) evaluated the performance of commercial branches of a large Canadian bank using DEA. Two production models were considered: one that looked directly at resource usage and was most useful to the branch manager and another that incorporated financial results and was geared more toward senior management. They concluded that an integrated performance measurement system can be built, encompassing several dimensions such as customer satisfaction, internal business, financial, and innovation and learning, and DEA can be incorporated in a general managerial framework for performance measurement.

Chen, Sun, and Peng (2005) studied the efficiency and productivity growth of commercial banks in Taiwan before and after the establishment of financial holding corporations. They used a DEA approach to generate

efficiency indexes and Malmquist productivity growth indexes for each bank. They found that member banks of financial holding corporations are more efficient than nonmember banks. Second, both member banks and nonmember banks of financial holding corporations demonstrated greater technical progress in the second time period during the year 2003.

Howland and Rowse (2006) assessed the efficiency of branches of a major Canadian bank by benchmarking them against the DEA model of US bank branch efficiency. They first used the model of bank branch efficiency by Golany and Storbeck (1999) to determine the appropriateness of their US model for Canadian banks. They then revised their model and, as part of model assembly, examined a correlation matrix of inputs and outputs to determine the relationships among potential output and input variables. The revised model had input variables and output variables that were similar but not identical to those used by Golany and Storbeck (1999). The revised model showed less variation in efficiency scores than the model used by Golany and Storbeck (1999) for replication.

Tsolas (2010) used a two-stage DEA model to analyze the performance of branches of a large commercial bank in Greece. The study disaggregated overall performance into profitability, efficiency, and effectiveness.

Lotfi et al. (2012) provided a framework to evaluate the performance of bank branches in terms of profitability efficiency and effectiveness using a two-stage DEA model. They applied this framework to bank branches of a large commercial bank in Iran.

Yavas and Fisher (2005) evaluated the operating performance of a bank branch relative to peer branches of the same bank in Southern California. Using data from 31 branches of a major bank, this study ranked bank branches on the basis of efficiency scores, identified areas of deficiency, and established the reference group against which a branch was evaluated. Of the 31 branches, 22 were found to need improvements in various areas.

Studies That Focus on Economies of Scale in the Banking Industry

Sufian (2007) used the DEA approach to evaluate trends in the efficiency of the Singapore banking sector and to distinguish among technical, pure technical, and scale efficiencies during the period of 1993–2003. The researchers found that the overall efficiency of the Singapore banking groups declined during the earlier part of the studies before increasing dramatically during the later period. The study also found that small Singapore banking groups tend to experience economies of scale (IRS), whereas large Singapore banking groups have experienced diseconomies of scale (DRS).

Rehman et al. (2015) used two-stage analysis to estimate the efficiency of the Pakistani banking sector from 1998 to 2009. In the first stage, they used DEA to estimate the efficiency of the banks. In step two, they regressed the estimated banking efficiency with macroeconomic indicators and corporate governance variables to determine the impact of economic conditions of a country and corporate governance practices on banking efficiency.

Yüksel, Mukhtarov, and Mammadov (2016) used a DEA model to compare the efficiency of Turkish and Azerbaijanian banks for the period of 2010–2014. They found Turkish banks were more efficient relative to Azerbaijanian banks.

Akeem and Moses (2014) performed an empirical analysis of the allocative efficiency of Nigerian commercial banks between 2002 and 2011 using a DEA model with three input variables (deposits, operating expenses, and assets) and four output variables (loan and advances, investment, interest income, and non-interest income). The study found that, to improve the allocative efficiency levels of banks across developing countries, it is imperative that banks use their labor and capital more efficiently, while simultaneously making enhanced and proficient use of technology in their operations. Furthermore, the study indicated that banks need to economize operating costs through internal restructuring, branch closures, and transacting business with customers profitably beyond the normal counter transactions of receiving cash and making payments.

Lin, Hsu, and Hsiao (2007) used DEA to study the relative efficiency of management and variations in managerial efficiency among 37 Taiwanese banks. They used DEA to estimate the relative managerial efficiency of each bank and found that all 37 banks showed an average efficiency score of 0.591. Their analysis also showed that the efficiency of 17 banks was declining.

Bergendahl and Lindblom (2008) developed principles to evaluate the efficiency of a savings bank using DEA as a method to consider the service orientation of savings banks. They determined the number of Swedish savings banks that were service efficient and the average degree of service efficiency in this industry.

Shanmugam and Das (2004) measured the technical efficiency of Indian banks in four ownership structures during the reform period of 1992–1999. They found that the efficiencies of raising interest margins were time invariant, whereas the efficiencies of raising other outputs (e.g., noninterest income, investments, and credits) were time varying.

Rezvanian, Rao, and Mehdian (2008) used a nonparametric frontier approach to examine the effects of ownership on the efficiency, efficiency change, technological progress, and productivity growth of the Indian banking industry over the period 1998–2003. They reported that foreign banks were significantly more efficient than Indian privately owned and publicly owned banks. The findings also provided evidence that a large number of banks operated below their optimal scale.

METHODOLOGY

Sample and Data Collection

The data for this study were obtained from financial statements published in the *Mergent Online* database, which provides company income statements and balance sheets. The study sample included 19 asset management and custody banks for the period of the fiscal year ending December 31, 2010 to December 31, 2017. To successfully apply the DEA methodology, we needed to consider homogenous units that perform similar tasks and accomplish similar objectives, and so our study included only companies identified by Standard and Poor NetAdvantage to be competitors. Exhibit 1 shows descriptive statistics of the variables used in this study.

EXHIBIT 1
Summary of the Data Used for 2010–2017

Statistics	Total Distribution Expenses ($)	General and Administrative Expense ($)	Total Expenses ($)	Total Assets ($)	Total Revenue ($)
2010					
Mean	5,535,412	26,835,360	230,794,040	1,059,195,731	184,033,191
Standard Deviation	16,389,721	106,726,576	948,702,325	4,309,296,608	714,665,490
Maximum	66,493,000	466,358,000	4,143,990,000	18,844,605,000	3,119,342,000
Minimum	2,497	2,965	8,807	131,192	3,961
2011					
Mean	5,650,535	32,354,830	191,733,866	1,225,990,821	192,700,243
Standard Deviation	16,880,681	129,608,904	775,518,646	5,010,775,772	745,290,874
Maximum	68,981,000	566,313,000	3,388,069,000	21,909,129,000	3,252,576,000
Minimum	2,698	3,042	8,979	133,986	4,537
2012					
Mean	5,662,151	33,224,107	195,111,471	1,692,396,246	244,518,577
Standard Deviation	15,806,186	128,995,083	766,576,789	6,800,020,950	945,167,974
Maximum	60,565,000	548,738,000	3,260,681,000	28,931,552,000	4,019,441,000
Minimum	3,036	2,945	9,229	134,729	4,692
2013					
Mean	4,866,477	27,452,928	216,093,241	1,648,011,920	365,467,552
Standard Deviation	13,535,113	108,565,004	885,102,616	6,790,032,013	1,514,176,038
Maximum	53,082,000	474,442,000	3,866,740,000	29,678,606,000	6,613,168,000
Minimum	3,236	3,095	9,721	144,576	5,253

(continued)

EXHIBIT 1 *(continued)*
Summary of the Data Used for 2010–2017

Statistics	Total Distribution Expenses ($)	General and Administrative Expense ($)	Total Expenses ($)	Total Assets ($)	Total Revenue ($)
2014					
Mean	4,991,304	32,373,185	215,510,075	1,745,412,807	410,693,924
Standard Deviation	12,935,835	125,846,498	882,594,342	7,209,942,032	1,713,994,254
Maximum	48,432,000	549,463,000	3,855,856,000	31,510,894,000	7,484,728,000
Minimum	3,236	3,095	9,721	148,810	12,296
2015					
Mean	38,923,078	5,244,298	184,634,244	1,387,233,841	277,299,007
Standard Deviation	148,608,154	19,151,999	748,927,389	5,448,565,403	1,125,937,529
Maximum	576,103,000	79,499,000	3,090,875,000	22,526,080,000	4,646,552,000
Minimum	3,276	3,082	10,028	145,342	12,200
2016					
Mean	37,651,577	183,619,671	6,158,984	1,712,173,695	324,270,018
Standard Deviation	138,919,695	731,988,254	12,637,453	6,585,168,376	1,280,423,067
Maximum	520,309,000	2,928,574,000	52,181,000	26,403,337,000	5,125,834,000
Minimum	138,919,695	731,988,254	12,637,453	6,585,168,376	1,280,423,067
2017					
Mean	33,799,105	233,583,293	11,356,562	2,217,513,104	449,102,782
Standard Deviation	124,488,365	931,685,900	32,533,568	8,590,445,446	1,778,658,277
Maximum	466,316,000	3,727,404,000	132,787,000	34,428,910,000	7,119,053,000
Minimum	3,399	3,051	9,813	147,470	12,075

Average total assets that these banks had under management rose from $1,059,195,731 in 2010 to $2,217,513,104 in 2017, an increase of more than 31 times (see Exhibit 1). Total revenue, on average, increased by more than 37 times over the same time period. Furthermore, from 2010 to 2017, there was a slight decline in average total expenses and a substantial increase in total distribution expenses and general administrative and selling expenses. All variables increased substantially across the board owing to the large increases in one of the 19 asset and custody banks, Blackstone Group LP.

Research Model

Charnes, Cooper, and Rhodes (1978) (CCR) proposed the original DEA methodology to measure the technical efficiency of organization units, called *decision-making units* (DMUs), assuming constant returns to scale (CRS). The objective of this technique is to measure DMU's effectiveness to use the available resources to produce outputs. We use the concept of efficiency or productivity (defined as a ratio of total outputs to total inputs) to assess the performance of DMUs. However, DEA calculates relative efficiency—that is, relative to the best-performing DMU or DMUs (in the case that multiple DMUs are the most efficient). DEA assigns the most efficient DMU an efficiency score of unity or 100% and allocates an efficiency score varying between 0% and 100% for the other DMUs relative to the best performance. Banker, Charnes, and Cooper (1984) (BCC) proposed a revised model as an extension of the CCR model to allow variable returns to scale (VRS). They showed that both CCR and BCC solutions allowed a decomposition of CCR efficiency into technical and scale components.

The generalized version of various DEA models that we use in this study is listed below:[2]

[2] There are multiple versions of the DEA model. For more details on the two-stage DEA model refer to Zhu (2003).

Frontier	Type	Input-Oriented	Output-Oriented
	CRS	$\text{Min } \theta - \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+)$ Subject to $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta x_{io}$ $i=1, 2, \ldots, m$ $\sum_{j=1}^{n} \lambda_j x_{jj} + s_i^+ = y_{iro}$ $r = 1, 2, \ldots, s;$ $\lambda_j \geq 0,\ j = 1, 2, \ldots n$	$\text{Max } \phi - \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+)$ $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io}$ $i = 1, 2, \ldots, m$ $\sum_{j=1}^{n} \lambda_j y_{rj} + s_r^+ = \phi y_{ro}$ $r = 1, 2, \ldots, s$ $\lambda_j \geq 0,\ j = 1, 2, \ldots n;$
	VRS: Add $\sum_{j=1}^{n} \lambda_j = 1$; NIRS: Add $\sum_{j=1}^{n} \lambda_j \leq 1$; NDRS: Add $\sum_{j=1}^{n} \lambda_j \geq 1$		

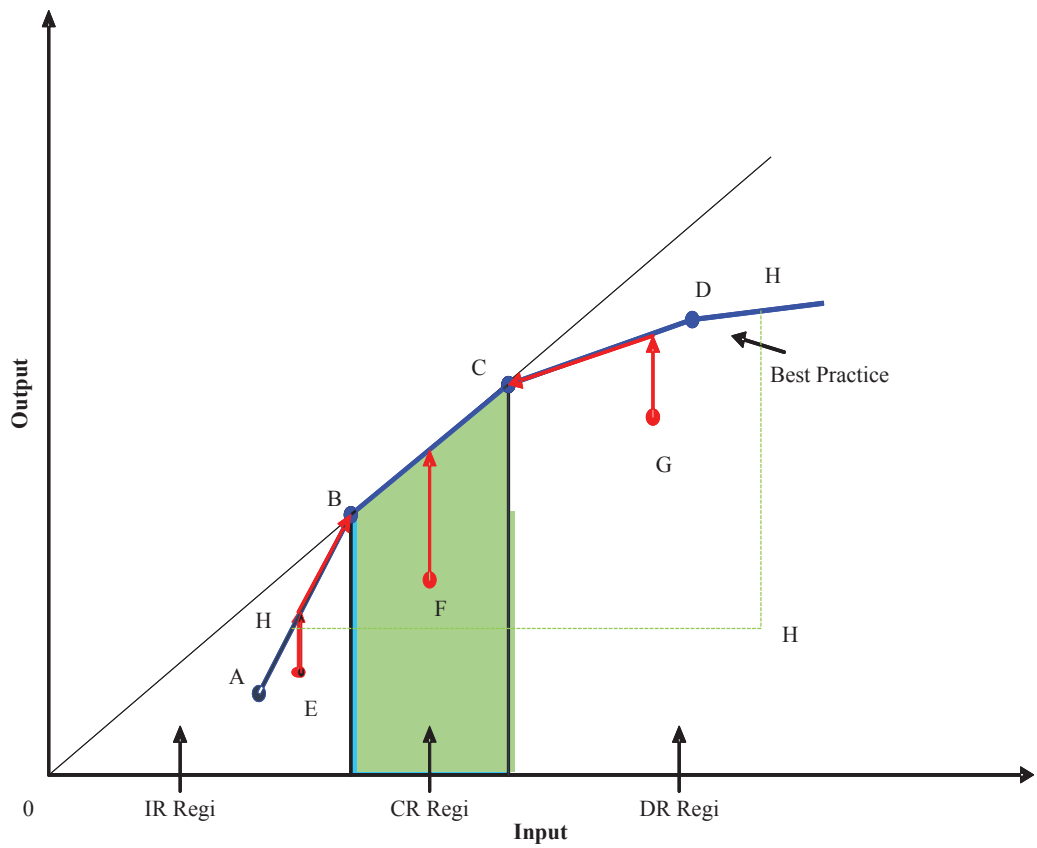
where s are the slack variables; x represent input variables; y represent output variables; λ is a scalar factor, and θ and ϕ represent the efficiency score of a DMU. In addition, NIRS is nonincreasing returns to scale, and NDRS is nondecreasing returns to scale.[3]

The fundamental DEA model assumes CRS, which means an increase in inputs leads to a proportional increase in outputs. However, the activities of smaller units can be significantly different when compared to larger units, indicating that VRS come into play. For these units, an increase in inputs does not imply a proportional change in the outputs. Thus, we use VRS, a type of frontier scale used in DEA, to estimate economies of scale when an increase or decrease in inputs or outputs does not result in a proportional change in the outputs or inputs, respectively (Cooper, Seiford, and Zhu 2011). The basic output of DEA, as explained in the preceding, is a collection of efficiency scores. The input efficiency scores represent the fraction of a firm's current costs used to produce a firm's output. Thus, an inefficient firm can achieve the same output with a lower cost. Alternately, using the output-oriented model, the efficiency score represents the proportion by which to increase the outputs using the same costs, done by following better practices. However, using scale efficiency, we compare two different DEA outputs: CRS efficiency and VRS efficiency. The objective is to investigate whether a firm is operating at the optimal size. If the firm is scale inefficient, we further investigate whether the firm is too large or too small, resulting in increasing or reducing the scale using existing best practices. Exhibit 2 illustrates the case of seven DMUs: A, B, C, D, E, F, and G.

Line OBC is the CRS frontier. AB, BC, and CD form the best practices envelopment or VRS frontier and exhibit IRS, CRS, and DRS, respectively. Line segment AB shows increasing RTS, BC displays CRS, and line segment CD exhibits DRS. Furthermore, the efficiency frontier encompasses all the DMUs (located below the envelope) that are inefficient, nonfrontier DMUs. The DEA model defines the magnitude of inefficiency as the distance from the envelope or frontier. However, corresponding to a VRS envelopment model being either input oriented or output oriented, there can be different efficient targets and RTS classifications. Zhu (2003) showed

[3] The original CCR model is CRS. It is built on the assumption that if an activity (x, y) is feasible, then, for every positive scalar t, the activity (tx, ty) is also feasible. Thus, we can scale the inputs and outputs linearly without increasing or decreasing efficiency. NDRS implies that changing all inputs by the same proportion changes the output by a greater extent than the proportional value. Similarly, in NIRS scaling up is interdicted, whereas scaling down is permitted. VRS combines these three possibilities—NDRS, CRS, and NIRS—for different ranges of output.

EXHIBIT 2
Returns to Scale Regions



Note: The CRS efficiency score is equal to the VRS efficiency score if and only if CRS prevail on DMU$_o$. $\Sigma_{j=1}^{n}\lambda_j^* < 1$ if and only if IRS prevail on DMU$_o$. $\Sigma_{j=1}^{n}\lambda_j^* > 1$ if and only if DRS prevail on DMU$_o$. in which λ_j^* is the alternate optima.

that we can explore RTS by calculating efficiency scores in empirical applications using the following theorem:

1. The CRS efficiency score is equal to the VRS efficiency score if and only if CRS prevail on DMU$_o$.
2. $\Sigma_{j=1}^{n}\lambda_j^* < 1$ if and only if IRS prevail on DMU$_o$.
3. $\Sigma_{j=1}^{n}\lambda_j^* > 1$ if and only if DRS prevail on DMU$_o$. *where* λ_j^* is the alternate optima.

Furthermore, using the concept of the most productive scale size (MPSS) (Banker, Charnes, and Cooper 1984), we can write a linear program for a given unique scale efficient target. As illustrated by Zhu (2003), consider the following input-oriented CRS envelopment model:

$$\text{Min } \Sigma_{j=1}^{n}\lambda_j^*$$

Subject to

$$\sum_{j=1}^{n}\lambda_j x_{ij} \le \theta^* x_{io,}\ i = 1, \ldots, m$$

$$\sum_{j=1}^{n}\lambda_j y_{rj} \ge y_{ro},\ r = 1, 2, \ldots, s \qquad (1)$$

$$\lambda_j \ge 0$$

where θ^* is the input-oriented CRS efficiency score.

Based on the optimal solution for Equation 1 (i.e., $\Sigma_{j=1}^{n}\lambda_j^*$), we can find the scale-efficient, largest MPSS target value for a DMU$_o$ using the following values:

$$\text{MPSS}_{\max}:\ x_{io}^T = \theta^* x_{io} / \sum_{j=1}^{n}\lambda_j^*$$

$$y_{io}^T = y_{ro} / \sum_{j=1}^{n}\lambda_j^* \qquad (2)$$

In addition to the computational details of a DEA model, we should consider many other factors while applying the DEA methodology. These factors relate to the choice of the DMUs for a given DEA application, selection of inputs and outputs, choice of a particular DEA model (e.g., CRS, VRS, IRS, and DRS) for a given application, and choice of an appropriate sensitivity analysis procedure (Ramanathan 2003). Furthermore, because DEA is a nonparametric technique, the DEA literature does not offer any standard specification search strategy. However, the quality of the analysis depends on the choice of the inputs and outputs included in the DEA model. The selection of DMUs depends on two main factors: homogeneity and the number of DMUs. The basic premise of homogeneity is a critical assumption. We should consider DMUs that are homogenous in every respect, perform similar tasks, and accomplish similar objectives.

Here, we can consider asset management and custody banks as homogenous because they compete with each other in the same industry. Another important consideration is the total number of DMUs used by the DEA model. The number of DMUs should be reasonable enough to span a wide arena of units, including high-performance units and underperforming units. In addition, a wide span of DMUs enables the DEA model to capture the relation between inputs and outputs. Some simple rules of thumb exist in the literature to guide the selection of inputs and outputs and the number of participating DMUs.[4] In this article, the relationship among these firms is an important consideration because they are all asset management and custody banks. Therefore, we select the envelopment models for our analysis. Furthermore, we should also choose whether to use input orientation or output orientation for the DEA model.

Typically, an analyst should choose an input orientation versus output orientation depending on whether there is more control over inputs versus outputs. Alternatively, the decision maker should determine whether the focus of the study is to maximize outputs (output orientation) or minimize inputs (input orientation). In our article, the focus of a bank is to minimize costs, so we consider an input-oriented model. Furthermore, the structure of the DEA model (in envelopment form) uses an equation (separate calculation) for every input and output. Therefore, we can use all the input and output variables simultaneously measured in their own units.

EMPIRICAL ANALYSIS

As illustrated, we use an input-oriented envelopment DEA model for our analysis. Thus, the objective of the DEA model is to minimize input for a given level of output. In addition, we use RTS and most productive scale size to study economies of scale among the given set of companies. Furthermore, we use the CRS and VRS efficiency models to investigate the RTS factor. The gap between the CRS and VRS efficiency frontiers will help identify the scale factor. Using the CRS model, we calculate CRS (total) efficiency, which comprises two parts: VRS (pure) efficiency and scale efficiency. Next, we compare the CRS and VRS efficiencies for a DMU to ascertain the nature of RTS.

To evaluate the optimal scale size for the asset management companies, we use one input variable and two output variables. The input variable is the total expenses incurred by a company. The output variables are total revenue and total assets. In addition, we use two different models to study the components of expenses: general and administrative expenses and distribution expenses. Because the distribution expenses information is missing for AMG, we only consider 18 companies for the distribution expenses model.

The study evaluates the cost efficiencies of asset management companies by tracking the scale efficiency gains (or losses) of each of the 19 companies. Using the DEA methodology, we can calculate an efficiency score for the 19 companies on a scale of 1 to 100 for VRS and CRS envelopment models. Furthermore, we use the

[4] The following are the guidelines for DMU model selection:

a. The number of DMUs is expected to be larger than the product of number of inputs and outputs (Avkiran 2001; Darrat, Topuz, and Yousef 2003) to discriminate effectively between efficient and inefficient DMUs. The sample size should be at least two or three times larger than the sum of the number of inputs and outputs (Ramanathan 2003).

b. The criteria for selection of inputs and outputs are also quite subjective. A DEA study should start with an exhaustive, mutual list of inputs and outputs that are considered relevant for the study. Screening inputs and outputs can be quite quantitative (e.g., statistical) or qualitative (simply judgmental, use expert advice, or use methods such as analytical hierarchy process) (Saaty 1980). Typically, inputs are the resources used by the DMUs or condition affecting the performance of DMUs. On the other hand, outputs are the benefits generated as a result of the operation of the DMUs and record higher performance in terms of efficiency. Typically, we should restrict the total number of inputs and outputs to a reasonable level. (Ramanathan 2003).

VRS efficiency score, CRS efficiency score, and λ_j^* (Exhibit 1) to calculate the RTS. In stage 1, we used total expenses as the input variable and total revenues and total assets as output variables. Exhibit 3 illustrates the CRS, VRS, and RTS efficiency of the 19 companies from 2010–2017.

In 2010, 13 of 19 companies show an increasing RTS, whereas two companies show DRS, and four companies show CRS (Exhibit 2). AMG, AMP, BEN, CNS, EV, FNGN, IVZ, JNS, LM, SEIC, TROW, VRTS, and WDR, which all manifest IRS, reaped the benefits of economies of scale—IRS. In the realm of IRS, the average input consumption declines with increases in output. Thus, these companies have not yet achieved the optimal size or the most productive scale size. Hence, a change in output for the IRS units leads to changes in input of less than one unit, leading to an overall reduction of the average input consumption. IRS for 13 asset management banks implies that there is room for further growth for these banks, and the operating cost will increase less than proportionately, which results in cost efficiencies in management; this will add value to the company, and shareholders' equity will increase.

On the other hand, BLK, CLMS, NTRS, and STT show CRS, which implies that these DMUs are located on both the CRS and the VRS efficiency frontiers. These companies have achieved their optimal size or efficient scale, and the scale does not affect their level of productivity. Typically, companies operating under CRS have minimized the consumption of average inputs with no impact of variation in outputs. Thus, an increase in output by one unit requires a proportionate increase in inputs by one unit also.

On the other hand, BK and BX are facing DRS. They have grown in terms of total assets under management, and the scale (or size) is negatively affecting the productivity of these DRS companies, leading to diseconomies of scale. The average consumption of inputs rises for these companies with the rise in output. For a company facing diseconomies of scale, a change in output by one unit results in a change in input by more than one unit. The disproportionate change in consumption occurs because these units are above their optimal size or most efficient scale. Thus, to improve their scale efficiency, BK and BX should reduce their size to achieve the most optimal size or rediscover their business processes so that their operating expenses decline relative to total assets under management.

Another way of interpreting the efficiency scores is a case-by-case analysis done by comparing the VRS and CRS scores. For instance, for 2010, companies BLK, CLMS, and NTRS are located on both the CRS and VRS frontiers. Thus, these companies do not have any inefficiency and are just facing CRS. Changes in output, therefore, are directly proportionate to changes in input. AMP is located on the VRS frontier but not on the CRS frontier. Thus, the source of inefficiency is inappropriate scale with IRS. A change in output of AMP by 1% will result in a change in input of less than 1%. Under this scenario, increasing the outputs to reach the most productive scale is the best option. Companies BK and BX are on the VRS efficiency frontier but not on the CRS efficiency frontier. Again, the inefficiency is due to an inappropriate scale. However, the source of efficiency in these instances is decreasing returns of scale. Thus, a variation in output of 1% would result in variation in input of more than 1%. The recommendation for these companies is a reduction in outputs to attain the most optimal size or a recalibration of internal business processes to reduce costs. AMG is located neither on the CRS nor on the VRS frontier. For AMG, the source of inefficiency is inappropriate scale and poor management. AMG is exhibiting IRS with a recommendation for expansion to the most productive scale. We do not have any example of this case in 2010, but in 2011 AMG, among others, is located on neither the CRS nor the VRS frontier. Again, in this case, the inefficiency stems from inappropriate scale and poor management. However, AMG is now facing the DRS and reduction in outputs that would enable it to achieve the most optimal scale.

Exhibit 3 also shows that as companies expand in terms of total assets and total revenues, they reach a point when size is too large and cost inefficiencies manifest in the form of DRS. In 2017, with the booming economy and its consequent impact on wealth management, only one asset management bank, BX, is on CRS. All of the other asset management banks show IRS, which means that an increase in size will disproportionately increase cost. Artificial intelligence–based portfolios, algorithmic trading platforms, back office automation, and other technological advances may contribute to reduced operating costs and resultant IRS that will add more value to shareholders' equity. In 2011, two companies (NTRS and SEIC) show CRS, six companies show DRS, and 11 companies show IRS. In 2012, two companies (BLK

Exhibit 3

CRS and VRS Efficiencies and RTS for Asset Management and Custody Banks for 2010–2017 Using DEA Model with General Administrative Expenses as Input Variable and Total Revenues and Total Assets as Output Variables

	2017				2016				2015				2014				2013				2012				2011				2010			
Company	CRS Efficiency %	VRS Efficiency %	Scale Efficiency %	RTS	CRS Efficiency %	VRS Efficiency %	Scale Efficiency %	RTS	CRS Efficiency %	VRS Efficiency %	Scale Efficiency %	RTS	CRS Efficiency %	VRS Efficiency %	Scale Efficiency %	RTS	CRS Efficiency %	VRS Efficiency %	Scale Efficiency %	RTS	CRS Efficiency %	VRS Efficiency %	Scale Efficiency %	RTS	CRS Efficiency %	VRS Efficiency %	Scale Efficiency %	RTS	CRS Efficiency %	VRS Efficiency %	Scale Efficiency %	RTS
AMG	3	4	82	I	2	2	69.1	I	12.6	12.6	99.9	D	68	68	99.8	I	75	76	99.5	I	71	71	99.7	I	19	41	45.1	D	69	69	99.3	I
AMP	6	100	6	I	3	100	2.73	I	38.76	100.00	38.76	I	66	100	65.6	I	43	100	42.8	I	42	100	42.3	I	43	100	42.8	I	44	100	44.4	I
BEN	3	3	93	I	2	2	87.11	I	12.21	75.60	16.15	D	100	100	100.0	C	89	89	99.9	I	86	87	98.5	D	19	84	23.0	D	80	80	99.8	I
BK	13	13	99	I	6	6	98.55	I	97.51	100.00	97.51	D	88	100	87.8	D	91	100	91.3	D	83	100	83.2	D	66	100	66.0	D	69	100	68.5	D
BLK	12	12	99	I	7	7	97.81	I	89.41	100.00	89.41	D	100	100	100.0	C	100	100	100.0	C	100	100	100.0	C	92	100	92.4	D	100	100	100.0	C
BX	100	100	100	C	100	100	100	C	19.49	100.00	19.49	D	90	100	90.4	D	100	100	100.0	C	76	100	75.7	D	24	100	24.0	D	47	100	46.7	D
CLMS													60	68	88.6	D	82	82	100.0	I	97	100	96.5	D	19	100	19.5	D	100	100	100.0	C
CNS	3	7	42	I	2	6	26.29	I	12.35	16.34	75.60	I	73	75	97.1	I	91	94	96.3	I	94	98	96.6	I	17	22	80.5	I	86	91	94.6	I
EV	3	4	75	I	1	3	57.80	I	10.58	11.01	96.08	I	70	70	99.4	I	85	86	99.2	I	82	82	99.5	I	17	37	46.7	D	83	84	99.1	I
FNGN													57	59	97.1	I	72	75	95.5	I	73	76	95.0	I	14	19	74.9	I	69	76	91.5	I
IVZ	3	3	93	I	1	2	83.82	I	17.81	49.48	36.00	D	62	62	99.9	I	70	70	99.8	I	73	73	99.5	D	24	63	39.0	D	68	68	99.7	I
JNS									100.00	100.00	100.00	C	66	66	99.3	I	80	82	98.7	I	82	83	99.1	I	21	22	98.3	D	85	86	99.0	I
LM	2	2	86	I	1	2	74.65	I	7.71	7.71	100.00	C	54	54	99.8	I	66	66	99.6	I	53	54	99.9	I	19	53	34.9	D	68	68	99.6	I
NTRS	14	14	98	I	7	7	96.16	I	95.17	97.48	97.62	D	100	100	100.0	C	99	99	99.9	I	98	98	99.9	I	100	100	100.0	C	100	100	100.0	C
SEIC	3	3	75	I	1	2	58.84	I	10.33	10.80	95.67	I	62	62	99.3	I	75	76	99.0	I	78	78	99.3	I	100	100	100.0	C	81	82	98.9	I
STT	11	11	99	I	6	6	98.01	I	81.45	87.38	93.22	D	100	100	99.9	D	89	92	97.7	D	90	94	95.0	D	93	100	92.8	D	82	82	100.0	C
TROW	3	4	90	I	2	2	81.16	I	13.79	53.26	25.90	D	85	85	99.9	I	97	98	99.7	I	100	100	100.0	C	18	72	25.3	D	95	96	99.6	I
VRTS	3	5	52	I	1	5	24.75	I	9.57	12.17	78.65	I	63	64	98.3	I	83	85	97.2	I	78	81	96.7	I	12	15	77.4	I	65	70	93.3	I
WDR	2	3	69	I	1	2	55.82	I	10.41	10.77	96.68	I	64	64	99.5	I	81	82	99.2	I	82	83	99.5	I	15	29	50.2	D	80	81	99.0	I

EXHIBIT 4

Target MPSS Values for Asset Management and Custody Banks for 2010 Using CRS Model with Total Expenses as Input Variable and Total Revenue and Total Assets as Output Variables

Company	CRS Efficiency	VRS Efficiency	RTS	Total Revenue	Total Assets	Total Expenses
AMG	69	69	IRS	$44,182,942.31	$172,120,614.06	$39,643,709.88
AMP	44	100	IRS	$1,123,445.20	$37,209,548.86	$2,497,900.00
BEN	80	80	IRS	$249,880,130.40	$457,156,822.65	$191,272,193.76
BK	69	100	DRS	$6,199,528.34	$122,686,376.49	$5,879,261.77
BLK	100	100	CRS	$8,612,000.00	$178,459,000.00	$5,614,000.00
BX	47	100	DRS	$16,473,763.75	$99,521,492.28	$21,885,100.21
CLMS	100	100	CRS	$326,039,000.00	$589,246,000.00	$199,538,000.00
CNS	86	91	IRS	$279,970,623.16	$422,971,434.33	$199,538,000.00
EV	83	84	IRS	$272,155,968.72	$310,722,079.69	$199,538,000.00
FNGN	69	76	IRS	$177,344,337.60	$345,282,945.39	$156,922,416.47
IVZ	68	68	IRS	$24,820,420.61	$145,491,630.86	$22,730,999.07
JNS	85	86	IRS	$102,194,386.82	$274,356,260.68	$73,861,277.31
LM	68	68	IRS	$57,266,964.86	$187,212,044.71	$52,107,919.82
NTRS	100	100	CRS	$2,729,000.00	$83,843,900.00	$2,497,900.00
SEIC	81	82	IRS	$263,062,034.40	$402,176,962.71	$199,538,000.00
STT	82	82	IRS	$8,242,598.15	$147,769,263.45	$6,497,962.44
TROW	95	96	IRS	$311,172,698.72	$478,040,649.91	$199,538,000.00
VRTS	65	70	IRS	$213,212,219.60	$219,635,607.18	$199,538,000.00
WDR	80	81	IRS	$262,450,058.38	$245,381,642.94	$199,538,000.00

and TROW) show CRS, six companies show DRS, and 11 companies show IRS. In 2013, seven banks show CRS, two companies show DRS, two companies show CRS, and 15 companies show IRS. In 2014, three companies (BLK, BEN, and NTRS) show CRS, four companies (BK, BX, CLMS, STT) show DRS, and 12 companies show IRS. In 2015, only six asset management banks show IRS, two show CRS, and nine show DRS. At the end of 2016, 15 out of 16 asset management banks in the sample show IRS, and only one bank, BX, shows CRS. Thus, most of the companies had decreasing economies of scale in only one year (2015). For all other years, we see most companies manifest IRS. Whenever companies reach their most productive scale size and begin to experience DRS, they make structural changes by adopting new technologies or changes in processes so that their total operating expenses decline and they can grow again without a proportionate increase in operating costs. Exhibit 3 also shows the trend in RTS for companies over the time period ranging from 2010 to 2017.

BLK, BEN, and NTRS adjusted their scale to reach the most productive scale exhibiting CRS. AMG, AMP, CNS, EV, FNGN, IVZ, JNS, LM, SEIC, TROW, VRTS, and WDR continue to work under IRS, with some firms intermittently working under DRS and CRS during some years. Likewise, BK, BX, CLMS, and STT continue to exhibit DRS, with some firms intermittently working under IRS and CRS intermittently during some years.

Exhibits 4 and 5 display the target MPSS for the 19 investment companies for the years 2010 and 2017 only.

Exhibit 4 illustrates the MPSS target of the input variable—total expenses and output variables—total revenues and total assets for the year 2010. In 2010, for AMG, the most productive scale size is $44,182,942.31 (total revenue), $172,120,614.06 (total assets), and $39,643,709.88 (total expenses) as the largest MPSS because AMG exhibits IRS. Similarly, BK exhibits DRS and therefore has the smallest MPSS at $6,199,528.34 (total revenue), $122,686,376.49 (total assets), and $5,879,261.76 (total expenses). Similarly, BLK exhibits CRS and has constant productive scale size MPSS at $8,612,000 (total revenue), $178,459,000 (total assets), and $5,614,000 (total expenses) that correspond to the current levels.

EXHIBIT 5

Target MPSS Values for Asset Management and Custody Banks for 2017 Using CRS Model and Total Expenses as Input

Company	CRS Efficiency	Total Revenue (MPSS)	Total Assets (MPSS)	Total Expenses (MPSS)
AMG	3	$16,906.39	$5,226.22	$13,531.74
AMP	6	$13,531.74	$13,531.74	$13,531.74
BEN	3	$17,029.02	$3,824.76	$13,531.74
BK	13	$30,550.66	$76,445.55	$33,859.88
BLK	12	$27,467.91	$39,651.79	$19,533.99
BX	100	$589,569.61	$233,463.82	$13,531.74
CNS	3	$18,612.58	$1,652.69	$13,531.74
EV	3	$16,070.98	$2,006.04	$13,531.74
IVZ	3	$14,613.48	$7,343.34	$13,531.74
LM	2	$12,261.98	$2,606.53	$13,531.74
NTRS	14	$33,106.49	$95,205.48	$38,913.71
SEIC	3	$14,861.14	$1,477.36	$13,531.74
STT	11	$25,962.75	$45,376.68	$23,650.26
TROW	3	$19,636.36	$2,527.57	$13,531.74
VRTS	3	$12,733.12	$6,346.63	$13,531.74
WDR	2	$13,892.21	$1,360.87	$13,531.74

Exhibit 5 shows most productive scale size for the year 2017.[5] In 2017, all asset management banks show IRS except BX, which is the only bank that shows CRS. Thus, for AMG, the most productive scale size is $16,906.39 (total revenue), $5,226.22 (total assets), and $13,531.74 (total expenses) as the largest MPSS because AMG exhibits IRS. Similarly, because BX exhibits CRS, it has the smallest MPSS at $589,569.6 (total revenue), $233,463.82 (total assets), and $13,531.74 (total expenses). Likewise, we can interpret the largest or smallest MPSS for the remaining companies depending on whether they show IRS or DRS.

Exhibit 6 illustrates a summary of the CRS and variable RTS relative efficiencies of 19 asset management and custody banks (17 banks for 2015 and 2016 and 16 banks for 2017) for the 2010 to 2017 time period using general and administrative expenses as input and total assets and total revenue as output.

For 2010, two companies show CRS, five companies show DRS, and 12 companies show IRS. In 2017, two companies have CRS, six companies have DRS, and eight companies show IRS. Because a majority of companies have increasing economies of scale across all years, we can infer that general and administrative expenses are an important factor in the overall cost structure of asset management and custody banks. In addition, BEN operated under CRS from 2010 to 2014, which means it maintains its most productive scale size; in 2015, BEN reports IRS for one year only. In 2016 and 2017, BEN shows DRS, which means that it exceeded its most productive scale size. BLK adjusted its scale to reach the most productive scale exhibiting CRS in 2016, but in 2017, it again reports DRS, which means that the most productive scale size has been exceeded. BLK needs to either scale back the size or recalibrate its business processes to cut costs. AMG, AMP, CNS, EV, LM, SEIC, VRTS, and WDR continue to work under IRS throughout the sample period. Likewise, BK, BX, CLMS, and STT continue to exhibit DRS with some firms intermittently working under IRS and CRS intermittently during some years. Overall, these companies exhibit RTS similar to that with the previous model using total expenses as input. In addition, NTRS and TROW seem to have adjusted to CRS using total expenses as input.

Exhibit 7 illustrates a summary of the CRS and variable RTS relative efficiencies of 19 asset management banks for the 2010 to 2017 time period (16 banks for 2015 and 2016 and 15 banks for 2017) using total distribution expenses as input and total assets and total revenue as output.

In 2017, we have three banks with CRS, three banks with DRS, and nine banks with IRS. For 2010, two banks have CRS, one bank has DRS, and 16 banks have IRS. Because the majority of companies have increasing economies of scale across all years, this implies that distribution expense is also a major contributor to cost efficiencies or inefficiencies. In addition, BK has been operating under CRS for total distribution expenses (and thus maintained its most productive scale size) throughout the sample period, except for 2015 and 2017, when it operated with DRS. AMG, CLMS, CNS, FNGN, JNS, NTRS, TROW, VRTS, and WDR continue to work under IRS for total distribution expenses throughout the sample period. Overall, these companies exhibit IRS, as opposed to the previous models using total expenses or general administrative expenses as input.

[5] There are 16 asset management and custody banks for the year 2017.

EXHIBIT 6

CRS and VRS Efficiencies and RTS for Asset Management and Custody Banks for 2010–2017 Using DEA Model with General Administrative Expenses as Input Variable and Total Revenues and Total Assets as Output Variables

	2010				2011				2012				2013				2014				2015				2016				2017			
Company	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS
AMG	20	20	98	I	21	21	98	I	22	22	98.33	I	23	23	98.61	I	21	68	30.18	I	10	10	93.67	I	28	29	99.08	I	31	32	99.14	I
AMP	28	100	28	I	27	100	27	I	30	100	29.97	I	30	100	30.11	I	35	100	35.33	I	17	100	16.64	I	33	100	33.19	I	35	100	35.12	I
BEN	100	100	100	C	100	100	100	C	100	100	100.00	C	100	100	100.00	C	100	100	100.00	C	39	40	97.93	I	87	100	87.36	D	91	100	91.17	D
BK	55	82	67	D	55	96	57	D	72	100	72.39	D	72	100	72.49	D	78	100	78.30	D	46	46	99.82	I	65	100	64.74	D	68	100	67.83	D
BLK	90	100	90	D	83	100	83	D	100	100	100.00	C	97	100	97.02	D	100	100	100.00	C	58	58	99.68	I	100	100	100.00	C	97	100	96.73	D
BX	42	100	42	D	36	100	36	D	52	100	51.82	D	81	100	81.31	D	63	100	62.56	D	100	100	100.00	C	10	100	9.83	D	11	100	10.58	D
CLMS	37	100	37	D	35	100	35	D	36	100	35.69	D	35	52	66.97	D	16	68	24.24	D	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CNS	23	32	73	I	26	34	77	I	30	37	82.01	I	29	35	82.98	I	20	75	26.92	I	11	17	66.18	I	28	32	86.18	I	29	34	86.68	I
EV	28	29	95	I	30	31	95	I	29	30	95.89	I	32	33	96.27	I	23	70	32.71	I	17	19	89.32	I	41	42	96.99	I	38	40	97.27	I
FNGN	39	62	63	I	41	62	67	I	53	70	75.35	I	50	63	79.53	I	39	59	66.98	I	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
IVZ	73	73	100	I	71	71	100	I	74	74	99.61	I	66	66	99.71	I	58	62	93.79	I	13	14	97.18	I	41	41	99.98	I	43	43	99.14	D
JNS	37	38	96	I	38	40	96	I	40	42	95.81	I	43	45	95.91	I	33	66	49.53	I	90	95	94.07	I	NA	NA	NA	NA	NA	NA	NA	NA
LM	26	26	99	I	26	27	99	I	26	26	99.02	I	25	26	98.91	I	22	54	40.10	I	14	15	94.07	I	42	42	99.49	I	45	46	99.53	I
NTRS	100	100	100	C	100	100	100	C	100	100	100.00	C	100	100	100.00	C	97	100	96.77	I	62	62	99.38	I	100	100	100.00	C	100	100	100.00	C
SEIC	32	35	94	I	35	38	94	I	41	43	94.87	I	44	46	95.39	I	33	62	52.25	I	8	9	88.82	I	19	20	97.17	I	19	19	97.26	I
STT	58	66	88	D	69	100	69	D	70	77	91.15	D	72	81	88.77	D	83	100	82.79	D	44	44	99.71	I	73	84	86.19	D	74	84	87.68	D
TROW	29	30	98	I	32	33	98	I	39	39	98.83	I	42	42	98.98	I	28	85	33.33	I	45	47	96.16	I	100	100	100.00	C	100	100	100.00	C
VRTS	37	55	68	I	33	45	74	I	36	44	82.37	I	36	42	86.58	I	30	64	47.39	I	10	15	69.46	I	29	33	88.81	I	37	40	93.51	I
WDR	62	66	95	I	58	61	95	I	68	71	95.75	I	72	74	96.31	I	47	64	72.33	I	25	27	90.03	I	60	62	96.63	I	45	46	96.03	I

Exhibit 7

CRS and VRS Efficiencies and RTS for Asset Management and Custody Banks for 2010–2017 Using DEA Model with Distribution Expenses as Input Variable and Total Revenues and Total Assets as Output Variables

	2010				2011				2012				2013				2014				2015				2016				2017			
Company	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS	CRS Efficiency	VRS Efficiency	Scale Efficiency	RTS
AMG	3	100	3.4	I	3	100	2.8	I	2	100	1.6	I	1	100	1.5	I	4	100	4.5	I	7	100	7.12	I	12	100	12.3	I	11	100	11.39	I
AMP	2	2	95.2	I	2	2	95.5	I	2	2	98.7	I	1	1	90.9	I	2	2	93.8	I	4	8	47.23	D	10	10	100.0	I	9	10	95.64	D
BEN	37	43	86.4	D	33	43	76.9	D	27	52	52.9	D	27	46	60.0	D	51	51	99.2	I	100	100	100.00	C	100	100	100.0	C	100	100	100.00	C
BK	100	100	100.0	C	100	100	100.0	C	100	100	100.0	C	100	100	100.0	C	19	19	98.6	I	66	87	76.28	D	87	87	100.0	C	83	90	92.92	D
BLK	100	100	100.0	C	100	100	100.0	C	90	100	90.3	D	100	100	100.0	C	100	100	100.0	C	12	100	11.99	D	33	100	33.0	D	47	100	47.47	D
BX	4	4	99.9	I	4	4	99.9	I	4	5	90.6	D	2	2	99.7	I	2	2	99.7	I	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CLMS	7	18	38.5	I	8	19	41.1	I	5	12	41.7	I	3	11	27.2	I	4	13	28.6	I	12	20	62.11	I	30	36	81.9	I	29	36	80.57	I
CNS	7	9	79.3	I	8	10	78.8	I	7	8	77.9	I	4	6	63.0	I	2	4	65.0	I	10	10	92.85	I	38	40	95.9	I	35	37	95.78	I
EV	4	13	27.6	I	4	12	29.8	I	4	11	32.4	I	2	10	23.1	I	2	9	26.5	I	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
FNGN	4	4	92.3	I	4	5	92.3	I	4	4	95.0	I	2	2	85.4	I	1	1	86.8	I	5	8	60.37	D	11	11	100.0	I	11	11	99.95	I
IVZ	6	8	77.6	I	5	7	74.3	I	5	7	70.3	I	3	5	52.3	I	3	5	54.9	I	100	100	100.00	C	NA	NA	NA	NA	NA	NA	NA	NA
JNS	3	3	90.0	I	3	4	89.1	I	3	4	90.4	I	2	2	76.6	I	2	2	77.8	I	7	7	100.00	D	19	20	99.0	I	20	20	98.89	I
LM	12	12	97.7	I	10	10	98.0	I	7	7	95.7	I	6	6	95.3	I	19	19	97.2	I	19	19	98.20	I	24	24	98.7	I	23	24	98.72	C
NTRS	5	7	75.5	I	5	6	73.3	I	5	7	73.8	I	3	4	58.5	I	3	4	61.8	I	10	11	92.12	I	24	25	96.1	I	24	25	95.77	I
SEIC	17	17	99.7	I	16	20	84.2	D	11	14	81.3	D	11	13	85.8	D	22	22	98.9	I	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
STT	27	30	89.0	I	22	25	89.0	I	20	21	91.9	I	10	12	81.4	I	7	9	83.6	I	38	61	61.90	D	100	100	100.0	C	100	100	100.00	C
TROW	4	11	33.0	I	3	7	37.6	I	3	7	42.2	I	2	5	32.8	I	1	4	36.6	I	6	9	66.10	I	16	19	80.6	I	18	22	82.74	I
VRTS	2	2	78.1	I	2	2	77.9	I	2	2	77.3	I	1	1	63.2	I	1	1	67.1	I	3	3	93.91	I	6	6	95.4	I	6	6	93.91	I
WDR	3	100	3.4	I	3	100	2.8	I	2	100	1.6	I	1	100	1.5	I	4	100	4.5	I	7	100	7.12	I	12	100	12.3	I	11	100	11.39	I

CONCLUSIONS

Using DEA, a nonparametric, performance measurement technique that enables managers to perform an efficiency analysis of investment management services, we identified the nature of RTS that indicates whether a firm should increase or decrease its scale or size to minimize average cost. To do so, we used two basic DEA models: the CRS model to calculate the CRS technical efficiency and the VRS model to calculate the VRS technical efficiency. By comparing these two models, we were able to identify sources of inefficiency. As illustrated, we can have CRS, IRS, and DRS. We looked at the variations in the scale efficiency of selected asset management banks from 2010 to 2017 to assess their progress. We found that investment management services companies typically operate under IRS (14 of 19 investment management companies showed IRS in 2010, and 16 of 19 showed IRS in 2017). Furthermore, we used two other models with general and administrative expenses and distribution expenses as inputs, respectively. Again, we found that most firms operated under IRS. Two companies had attained their optimal scale size. Six companies showed DRS in 2017 and needed to find a way to reduce general and administrative expenses, which may be contributing to their inefficiencies. However, nine firms showed IRS while having distribution costs as input, which signified that the asset management industry has scope to expand revenues and assets while maintaining the current levels of distribution costs. Managers can use the benchmarking factor (efficiency score), peer analysis, and slack values to improve or adjust the inputs and outputs for their company. In addition, we calculated the largest or smallest MPSS for the remaining investment management services companies, depending on whether a company shows IRS or DRS using a scale efficiency target.

Our findings are consistent with previous studies with regard to economies of scale and mergers and acquisitions in the industry. Neal (2004) concluded that diseconomies of scale set in very early and hence are not sufficient to allow mergers between large banks to proceed. However, we also show that many companies with DRS were able to achieve CRS and sometimes even IRS by finessing their business processes and business models.

This study also illustrated the use of DEA as a performance measurement technique to enable decision makers to conduct an efficiency analysis and determine many target values by comparing their firm with others in the industry. The major strength of DEA is its objectivity: DEA identifies efficiency ratings based on numerical data as opposed to subjective human judgment and opinion. In addition, DEA can handle multiple inputs and outputs measured in different units, and unlike statistical methods of performance analysis, DEA is nonparametric and does not assume a functional form relating inputs and outputs. Thus, this article illustrates DEA as a viable business intelligence tool for decision making.

DEA enables decision makers to perform four main tasks: calculate efficiency scores indicating whether a firm is efficient or has scope for improvement; set target values for input and output to explain the levels by which inputs should be contracted or outputs should be expanded to become 100% efficient within the industry; identify the nature of RTS, indicating whether a firm should increase or decrease in size to attain the most productive scale; and identify a set of benchmarks or role models specifying a set of processes that should be adjusted to improve a firm's practices. Typically, DEA analysis benchmarks a firm against others in the industry to provide relative analysis. Thus, decision makers can conduct a comprehensive comparative analysis using DEA from time to time to ascertain a firm's standing within the industry.

Findings from our research are also likely to be of use to the capital market industry and regulatory agencies such as the Securities and Exchange Commission.

REFERENCES

Akeem, U., and F. Moses. 2014. "An Empirical Analysis of Allocative Efficiency of Nigerian Commercial Banks: A DEA Approach." *International Journal of Economics and Financial Issues* 4 (3): 465–475.

Avkiran, N. 2001. "Investigating Technical and Scale Efficiencies of Australian Universities through Data Envelopment Analysis." *Socio-Economic Planning Sciences* 35 (1): 57–80.

Balfour, R., S. Joo, H. I. H. Whited, and J. W. Lin. 2015. Assessing the Comparative Performance of Banking Branches. *Benchmarking* 22 (5): 963-972.

Banker, R., A. Charnes, and W. Cooper. 1984. "Some Models for the Estimation of Technical and Scale Inefficiencies in Data Envelopment Analysis." *Management Science* 30 (9): 1078–1092.

Bergendahl, G., and T. Lindblom. 2008. "Evaluating the Performance of Swedish Savings Banks According to Service Efficiency." *European Journal of Operational Research* 185 (3): 1663–1673.

Charnes, A., W. Cooper, and E. Rhodes. 1978. "Measuring Efficiency of Decision-Making Units." *European Journal of Operational Research* 2 (6): 429–444.

Chen, Y., L. Sun, and C. Peng. 2005. "Commercial Banks' Performance in Taiwan." *International Journal of Performance Management* 7 (4): 444–463.

Cooper, W. W., L. M. Seiford, and J. Zhu. Data Envelopment Analysis: History, Models, and Interpretations. In: Cooper, W. W., L. M. Seiford, and J. Zhu. Eds., *Handbook on Data Envelopment Analysis*. New York: Springer US, 1–39, 2011.

Darrat, A., C. Topuz, and T. Yousef. 2003. "Assessing Bank Efficiency of Banks in an Emerging Market: The Kuwaiti Experience in the 1990s." *Studies in Economics and Finance* 21 (2): 1–21.

Golany, B., and I. Storbeck. 1999. "A Data Envelopment Analysis of the Operational Efficiency of Bank Branches." *Interfaces* 29 (3): 14–26.

Halkos, G. E., and D. S. Salamouris. 2004. Efficiency Measurement of the Greek Commercial Banks with the Use of Financial Ratios: A Data Envelope Analysis Approach. *Management Accounting Research* 15 (2): 201–224.

Howland, M., and J. Rowse. 2006. "Measuring Bank Branch Efficiency Using Data Envelopment Analysis: Managerial and Implementation Issues." *INFOR* 44 (1): 49–63.

Kao, C., and S. Liu. 2004. "Predicting Bank Performance with Financial Forecasts: A Case of Taiwan Commercial Banks." *Journal of Banking & Finance* 28 (10): 2353–2368.

Lin, Y. H., H. Hsu, and C. K. Hsiao. 2007. "Measuring Efficiency of Domestic Banks in Taiwan: Application of Data Envelopment Analysis and Malmquist Index." *Applied Economics Letters* 14 (11): 821–827.

Lotfi, F., A. Eshlaghy, M. Shafiee, H. Saleh, H. Nikoomaram, and S. Seyedhoseini. 2012. "A New Two-Stage Data Envelopment Analysis (DEA) Model for Evaluating the Branch Performance of Banks." *African Journal of Business Management* 6 (24): 7230–7241.

Malhotra, D., R. Poteau, and A. Malhotra. 2014. "An Empirical Analysis of Cost Efficiencies in the Indian Banking Industry." *International Journal of Business* 19 (4): 361–380.

Neal, P. 2004. "X-Efficiency and Productivity Change in Australian Banking." *Australian Economic Papers* 43 (2): 174–191.

Paradi, J., and C. Schaffnit. 2004. "Commercial Branch Performance Evaluation and Results Communication in a Canadian Bank—A DEA Application." *European Journal of Operational Research* 156 (3): 719–734.

Pille, P., and J. Paradi. 2002. "Financial Performance Analysis of Ontario (Canada) Credit Unions: An Application of DEA in the Regulatory Environment." *European Journal of Operational Research* 139 (2): 339–350.

Ramanathan, R. *An Introduction to Data Envelopment Analysis—A Tool for Performance Measurement*. New Delhi, India: Sage Publications, 2003.

Rehman, R., J. Zhang, R. Ali, and A. Qadeer. 2015. "Does Growing Economy and Better Governance Impede Banking Efficiency? A DEA Analysis." *The Journal of Applied Business Research* 31 (6): 2213–2224.

Rezvanian, R., N. Rao, and S. M. Mehdian. 2008. "Efficiency Change, Technological Progress and Productivity Growth of Private, Public and Foreign Banks in India: Evidence From the Post-Liberalization Era." *Applied Financial Economics* 18: 701–713.

Saaty, T. L. *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.

Shanmugam, K. R., and A. Das. 2004. "Efficiency of Indian Commericial Banks During the Reform Period." *Applied Financial Economics* 14 (9): 681–686.

Sufian, F. 2007. "Trends in the Efficiency of Singapore's Commercial Banking Groups A Non-Stochastic Frontier DEA Window Analysis Approach." *International Journal of Productivity and Performance Management* 56 (2): 99–135.

Tsolas, I. 2010. "Modeling Bank Branch Profitability and Effectiveness by Means of DEA." *International Journal of Productivity and Performance Management* 59 (5): 432–451.

Yavas, B. F., and D. M. Fisher. 2005. "Performance Evaluation of Commercial Bank Branches Using Data Envelopment Analysis." *Journal of Business and Management* 11 (2): 89–102.

Yüksel, S., S. Mukhtarov, and E. Mammadov. 2016. "Comparing the Efficiency of Turkish and Azerbaijani Banks: An Application with Data Envelopment Analysis." *International Journal of Economics and Financial Issues* 6 (3): 1059–1067.

Zhu, J. *Quantitative Models for Performance Evaluation and Benchmarking.* Boston: Kluwer's International Series, 2003.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Do Health Care Mutual Funds Provide Healthy Risk-Adjusted Returns?

Srinidhi Kanuri and Davinder Malhotra
The Journal of Wealth Management
https://jwm.pm-research.com/content/22/4/90

ABSTRACT: *Specialty mutual funds, such as health care funds, offer an investor the opportunity to focus on high total return, income, and further diversification within a specific industry. Although extensive research has been conducted on mutual fund performance as well as other specialty mutual fund performance, little research has focused on the performance of health care mutual funds. This article seeks to decode the characteristics of health care mutual funds and their performance by examining their risk-adjusted rate of return from 2001 to 2018. The authors find that, on average, equally weighted portfolios of health care mutual funds did not perform better relative to Dow Jones Health Care Index. Net four-factor alphas of health care mutual funds were positive but insignificant. In addition, based on the two performance models studied, these categories of funds have not consistently created positive net alphas for their investors over longer time periods, such as measured the 2001–2018 period.*

Cost Efficiencies in the Management of Commodity Mutual Funds

D. K. Malhotra, Tim Mooney, and Raymond Poteau
The Journal of Wealth Management
https://jwm.pm-research.com/content/21/3/59

ABSTRACT: *This article examines the determinants of cost efficiencies in the U.S. commodity mutual fund industry from 2001 to 2016. Empirical results show that cost increases in the U.S. commodity mutual fund industry have been less than proportional to increases in fund assets, pointing to economies of scale for the industry. Average cost elasticity varies by fund size, existence of 12b-1 fees, load versus no-load funds, and institutional versus retail funds. Funds without a 12b-1 plan show larger economies of scale than funds with a 12b-1 plan. Institutional funds show greater economies of scale than retail funds since 2010.*

Evaluating the Performance and Diversification Benefits of Emerging-Market Exchange-Traded Funds

Srinidhi Kanuri, Davinder Malhotra, and James Malm
The Journal of Wealth Management
https://jwm.pm-research.com/content/20/4/85

ABSTRACT: *This study evaluates the performance and diversification benefits for U.S. investors of emerging-market exchange-traded funds (ETFs) since their inception in January 2003 through June 2015 by comparing their absolute and risk-adjusted performance with the iShares Core S&P 500 ETF (IVV). The authors find that the emerging-market ETF portfolio has very low correlations with IVV during the period of the study. Emerging-market ETF portfolios delivered better absolute performance (returns and wealth) but also had much higher risk (standard deviation of returns). However, the risk-adjusted performance (Sharpe and Omega ratios) of IVV was better than that of the emerging-market ETF portfolio. They also look at the effect of adding some emerging-market ETFs to IVV during the period of study. The authors find that adding some emerging-market ETFs to IVV leads to higher absolute returns, better risk-adjusted performance (Sharpe and Omega ratios), higher cumulative returns, and increased wealth for U.S. investors. Results are statistically significant at 1% in all cases. Therefore, U.S. investors should add some emerging-market ETFs to their domestic allocation based on their risk tolerance for better performance (absolute and risk-adjusted).*

Submission Guidelines

PORTFOLIO MANAGEMENT RESEARCH

All submitted manuscripts must be original work that has not been published in another form such as another journal, magazine, or book chapter. An author may submit manuscripts for consideration by only one journal in the **Portfolio Management Research** (PMR) family at a time. Additionally, by submitting a manuscript for consideration by a PMR publication, an author agrees to refrain from submitting the manuscript elsewhere until an editorial decision has been made. If a manuscript is rejected, the author may submit it to one of the other journals in the PMR family.

PMR accepts submissions only through our online article submission system. To submit a manuscript, an author must upload it to the online system and complete the required checklist, including agreement to the terms of the PMR copyright agreement. We do not require any specific style or format for an initial submission. However, upon acceptance of an article for publication, the author will be expected to bring their manuscript into conformity with the stylistic and mechanical specifications described below. Also, at that time, the author will be expected to manually sign the PMR copyright agreement and to remove all prior versions of the article from everywhere that they have been posted, including SSRN and all other websites.

Target Length: The target length for articles is 4,000 words, not including the abstract, exhibits, and references. However, the editor of a journal may accept articles that are substantially longer or shorter. Articles shorter than 2,500 words or longer than 7,500 words are rarely accepted.

Title: You should give your article a short descriptive title. In general, an article's title should be no more than 12 words. You should try to avoid including words that carry little or no meaning.

Abstract: You should write a strong abstract for an article appearing in a PMR publication. The target length for abstracts is 160 words. The abstract should be non-technical, should not contain any references, and should define any acronyms on first use. An article's abstract is the only part of the article that appears on the unrestricted portion of the related journal's website. In the printed version of a journal, the abstract appears in the table of contents (as if it is written by the editorial team, not the authors). It does not appear directly before the article's main text.

Introductory Material: Please write strong introductory paragraphs. The introductory material is an important part of an article. It serves as a substitute for the abstract at the start of the article. Because an abstract is not presented right before its associated article, a reader will not necessarily have read the abstract before reading the main text. An article's introductory material should describe the article's main points and briefly explain the reasoning or analysis that supports them. The introductory material should explain why the findings, results, or conclusions are important and how they add to prior research in the area. If applicable, the introductory material should convey how an investment professional can use the findings, results, or conclusions in a practical way. In most cases it is helpful to include a "roadmap" paragraph at the end of the introductory material. The purpose of the roadmap paragraph is to inform readers about how the article is organized into different sections and what is covered in each one.

Headings: Articles in PMR publications can have up to three levels of headings. The first heading in an article should appear after the introductory material. Articles in PMR publications do not include any heading before an article's introductory material.

Conclusion: Please end the main text of an article with a short conclusion that briefly summarizes the article's main point. The conclusion may resemble the article's abstract, though it need not omit references and technical terms used in the article. The conclusion should not introduce new material.

Footnotes: PMR publications now use footnotes rather than endnotes. Footnotes generally should not be used for references but rather for peripheral or tangential commentary that would not be suitable for inclusion in the main text.

References: PMR follows the "author-date" reference format of *The Chicago Manual of Style* (CMS), 17th Edition, Chapter 15 both for references included in the text of an article and for the list of references at the end of an article. The list of references should immediately follow the main text.

Style, Usage, and Grammar: PMR follows CMS with respect to matters of style, usage, and grammar. This includes standards for hyphenation, abbreviations, and capitalization.

Exhibits: Please call all tabular or graphical material Exhibits. Number exhibits with Arabic numbers consecutively in order of appearance in the text. Do not call them Table 1 and Figure 1. An article's text should refer to all of its exhibits. In general, the text should clearly explain the point supported or demonstrated by each exhibit. Additionally, except for the simplest tables and charts, the text should explain how to read each exhibit. For tables that report statistical results, the text should direct readers to the key values and clearly explain what they mean. For an equation, the text should describe and explain the relationship represented by the equation and clearly define all terms used in the equation. In many cases, it is better to place content that is heavy with equations in an appendix. Within the PMR family of journals, *The Journal of Fixed Income* and *The Journal of Derivatives* generally allow for the highest proportions of quantitative material in the main text.

Exhibit Presentation: Authors are encouraged to submit articles with exhibits in color. Although PMR publications are generally printed in black and white, the online version can show color exhibits. Also, upon article acceptance, if you wish to have your exhibits printed in color, you may do so at a cost of $1,200. Please make sure that all categories in an exhibit can be distinguished from each other and match the categories in any legend connected with an exhibit. If you use color, please make sure that each exhibit will be legible if printed in black-and-white. This is important for all articles because a reader may print an article on a black-and-white printer.

Copyright Agreement: PMR's copyright agreement form must be signed prior to publication. Only one author's signature is necessary. Should your article not be accepted the copyright agreement will be considered null and void. Upon acceptance of the article, no further changes are allowed, except with the permission of the editor.

NEW JOURNAL

THE JOURNAL OF
impact&esg
INVESTING

Are you conducting research on how ESG factors are impacting the financial industry?

How are these new factors influencing investments?

What are the possible risks they pose?

What opportunities do they provide to make greater returns?

submit your article now

mestelle@hillcrestasset.com

Portfolio Management Research is looking for contributions to its newest journal, ***The Journal of Impact and ESG Investing.***

The journal will provide investment professionals with research across multiple asset classes, providing actionable solutions to the practical problems these new factors pose, while showcasing the opportunities they present to enhance and develop portfolios.

jesg.pm-research.com

PORTFOLIO
MANAGEMENT
RESEARCH

ACTIONABLE RESEARCH
SMARTER INVESTING

PORTFOLIO MANAGEMENT RESEARCH is the leading provider of independent financial research for all professionals in the investment management industry. Delve further than ever before into our content, on our new global platform.

UNRESTRICTED ACCESS

- 400 ANNUAL ARTICLES
- COVERING 60+ TOPICS
- AN ARCHIVE OF OVER 10,000 ARTICLES

PORTFOLIO
MANAGEMENT
RESEARCH

CONTACT US: (646) 931-9045 | pm-research.com@pageantmedia.com | www.pm-research.com